The Dissertation Committee for Sarah Emily Elias Stephens
certifies that this is the approved version of the following dissertation:

Longitudinal Predictions Using Alternative Binning to Reduce Regression to the Mean

Committee:

Michael Marder, Supervisor

Jill Marshall

Vernita Gordon

Austin Gleeson

Harry Swinney

Longitudinal Predictions Using Alternative Binning to Reduce Regression to the Mean

by

Sarah Emily Elias Stephens

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2017

Acknowledgements

I would like to express my deep gratitude to my advisor, Michael Marder, for his support, encouragement, understanding, and enthusiasm. It cannot be overstated how grateful I am to have an advisor who cares about my work, my life, and my success. He helped me accomplish the seemingly impossible in just three years. He created a balanced environment with just the right amount of independence, structure, guidance, creativity, and fun. He inspires me to be excited about my work and to use my skills to make the world a better place.

I would like to offer my special thanks to my committee members for their time and effort, but also for their support over the years, without which I would not have made it this far. Jill Marshall has shared her expertise in education research and she acts in many ways like a co-advisor. Austin Gleeson took me under his wing and mentored me when I first began teaching. Vernita Gordon has provided personal and academic advice on many occasions and I look up to her as a successful female physicist and role model. Harry Swinney does a tremendous job organizing the CNLD, which provides me with a supportive community.

I would like to acknowledge the help provided by my fellow graduate students. Matt Guthrie is an incredible co-worker and friend. He is extremely

reliable, hard-working, and giving. He dedicated a lot of time and mental energy to helping me out despite that endless to-do list of his. It is a comfort to know that he is always there with suggestions, sarcasm, and pizza. Dave McGhan and Tony Bendinelli paved the way for me and left me with a solid foundation to stand on, which was invaluable. Caitlin Hamrock and Ben David are my partners-in-crime at the ERC and they are amazing co-workers and friends. They all kept me sane and happy.

I would like to thank my family and my friends for providing support and encouragement. My parents push me to do better than my best and somehow they always know I can do it. My friends have made Austin my home and they provide endless love, community, and support. This was a huge chapter in my life and they were there through it all.

Most of all, I would like to thank my husband Ben Stephens. He is so supportive, encouraging, and inspiring. He dedicates his time and energy to helping me, even though he is living through the ups and downs of graduate school himself. I am excited for the future, despite the numerous unknowns, because he will be with me through it all. He is just the best thing ever.

Longitudinal Predictions Using Alternative Binning to Reduce Regression to the Mean

Sarah Emily Elias Stephens, Ph.D.
The University of Texas at Austin, 2017

Supervisor: Michael Marder

Educational policies in Texas that regulate the evaluations of students, teachers, and schools, can have profound impacts on the success of those individuals and institutions. The evaluations are largely based on the outcomes from standardized exams, as well as graduation rates and college preparedness. The analysis of standardized exam scores and policy impacts must be accurate, rapid, and reliable if it is to inform new policies. The possibility of using year by year longitudinal series of exams to extract predictions about policy interventions is greatly impacted, in practice, by a statistical phenomenon known as *regression to the mean*. I present a novel method, inspired by statistical and fluid mechanics, to address this problem, called Alternatively Binned Streamlines. I justify the use of this method through a simple theory. Then I apply it to the Texas State Longitudinal Data System, which contains standardized testing data for primary and secondary school students between 2003 and 2015. I show that regression to the mean can largely be eliminated, making

it possible to predict the longitudinal performance of aggregated students, using only two or three years of data, with acceptable accuracy. Through these predictions, I also identify the effects of a state-wide intervention called the Student Success Initiative. Thus, I demonstrate that Alternative Binning provides rapid analysis of policy impacts and predictions of longitudinal student performance with the ability to inform policy.

Table of Contents

Acknowledgements **iv**

Abstract **vi**

List of Tables **xi**

List of Figures **xiii**

Chapter 1. Introduction **1**

1.1 Standardized Testing in the United States 4

1.1.1 No Child Left Behind Act (NCLB) 4

1.1.1.1 School Accountability 7

1.1.1.2 Value-Added Modeling 9

1.2 Texas Setting 11

1.2.1 Education Research Center (ERC) 11

1.2.1.1 ERC Access: FERPA 13

1.2.2 Texas Standardized Exams 16

1.2.3 Student Success Initiative (SSI) 22

1.3 Testing Theory and Design 26

1.3.1 Classical Test Theory 27

1.3.2 Item Response Theory 29

1.3.3 TAKS Test Design 33

1.3.4 Scaled vs. Raw Scores 37

Chapter 2. Background **40**

2.1 Hierarchical Linear Models 42

2.1.1 Individual Growth Models 46

2.1.2 Structural Equation Models 47

2.1.3 Grouped Multilevel Models 49

2.2 Age-Period-Cohort Effects 50
2.2.1 Cross-Sectional Models 54
2.2.2 Single-Cohort Models 56
2.2.3 Accelerated Longitudinal Models 57
2.3 Foundation Techniques 62
2.3.1 Trajectory Plots 65
2.3.2 Streamline Plots 69

Chapter 3. Methods **76**
3.1 Regression to the Mean 76
3.2 Regression to the Mean in the Literature 87
3.3 Matrix Re-binning 89
3.4 Alternatively Binned (AB) Streamlines 92
3.4.1 Non-Traditional Students 101
3.5 Non-Linear Regression to the Mean 105
3.5.1 Quadratic Conditional Expectation Value 105
3.5.2 Normally Distributed Random Components 108

Chapter 4. Results **110**
4.1 AB Cohort Streamlines 110
4.1.1 Future Predictions 113
4.1.2 Sorting by Reading Score 116
4.2 Student Success Initiative 117
4.3 Demographic Differences 121
4.4 Physics Course Taking 124
4.5 Course Requirements 128
4.6 Teacher Certification 131

Chapter 5. Conclusions **138**

Appendix **141**
0.1 Z-score Properties 142
0.2 Pearson Correlation Coefficient Magnitude 142
0.3 Linear Conditional Expectation Value 143

Bibliography **145**

Vita **158**

List of Tables

1.1 Adequate Yearly Progress (AYP) standards in Texas between 2002 and 2014 [1]. This equals the percentage of students expected to reach proficiency. 7

1.2 TAKS exam subjects by grade [2]. 17

1.3 Timeline of Student Success Initiative (SSI) programs by school year [3]. 22

1.4 Total appropriated funding and impacted grades for the SSI by school year [3, 4]. The budget has decreased substantially in recent years. 23

1.5 Lower bounds for the reliability on the TAKS mathematics exams in 2010 [2]. 35

2.1 Results from a two-level longitudinal HLM measuring natural science knowledge. Replicated from Raudenbush and Bryk (2002), p.165 (adjusted notation). *se* is standard error, *df* is degrees of freedom, *OLS* is ordinary least squares. 45

2.2 Grade-Period table: each grade is represented by a row, each period (school year) is represented by a column, and each cohort is represented by a diagonal (highlighting the cohort of students that graduated in 2012). 52

2.3 Example of data used in an ALD with four cohorts, covering ten grades in only four years. 58

2.4 Summary of the differences between trajectories and snapshot streamlines. 73

3.1 Results from regression of the z-score in 4th grade with respect to the z-score in 3rd grade. Squared and cubic terms were significant but small. By adding quadratic and cubic terms, the R^2 value changed from 0.514 to 0.515. 82

3.2 Anticipated sequences of scores for trajectories and streamlines with respect to the initial score. 87

3.3 The grades, cohorts (by graduation year), and school years of the data used for the AB snapshot streamline of 2003-2005. 95

3.4 Comparison between the score sequences within the trajectory, cohort streamline, and AB cohort streamline frameworks, assuming each pair of exams has the same correlation coefficient. 100

3.5 Comparison between the score sequences within the trajectory, cohort streamline, and AB cohort streamline frameworks, using the Pearson correlation coefficients computed from the data for the cohort of 2012. 100

3.6 Sequences of anticipated z-scores for trajectories and AB cohort streamlines for the cohort of 2012. The first and third columns show the scores with respect to the 3rd grade score. The second and fourth columns show the scores with respect to the 4th grade score. 101

4.1 Results from regression of the mathematics z-score in 4th grade with respect to the reading z-score in the same grade. By adding quadratic and cubic terms, the R^2 value changed from 0.4127 to 0.4199. 116

4.2 Demographic breakdown of the 9th graders in 2008-2009. The traditional cohort limits the total to the population of students who remained in the cohort, graduating in 2012. 122

4.3 Course taking percentages for the traditional cohort of 2012. . . . 125

4.4 The odds ratios for taking AP physics, which equal the exponentiated coefficients from the logistic regression in Equation 4.1. . . . 128

4.5 Mathematics and science course requirements in Texas between 2001 and the present [5]. 130

4.6 Percentages of teachers teaching the 2012-2015 cohorts who have a STEM certification. Lower level courses have a smaller proportion of STEM certified teachers. 132

List of Figures

1.1 College readiness in Texas schools by poverty quartile. The drop in college readiness by 2015 is likely due to the exam requirement changes since HB-5. Figure provided by Marder. 21

2.1 Replicated path diagram [6] corresponding to Equation 2.11. . . . 48

2.2 Flow plots for low-income students between 2003 and 2007 representing score changes for each grade and score bin. The tan band surpassed the Met Standard cut-off score, the green band achieved Commended Performance [7]. 64

2.3 Trajectories for the cohort of students who graduated in 2012, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group. 66

2.4 Trajectories for the cohort of students who graduated in 2011, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group. 67

2.5 Trajectories for the cohort of students who graduated in 2013, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group. 68

2.6 Arrow plot and corresponding streamlines for the cohort of 2012. The arrows show the change in score by grade and score bin. The streamlines show interpolated scores over time based on the arrows. 71

2.7 Arrow plot and corresponding streamlines for each grade transition between 2003 and 2004. The arrows show the change in score by grade and score bin. The streamlines show interpolated scores through the grades based on the arrows. 72

2.8 Cohort streamlines and trajectories for the cohort of 2012. The convergence of the streamlines is due to regression to the mean. . . 74

2.9 Snapshot streamlines for 2003-2004 and trajectories for the cohort of 2012. The convergence of the streamlines is due to regression to the mean. 75

3.1 Distributions of mathematics TAKS percent scores for the cohort of students graduating in 2012 in 3rd, 4th, and 5th grade. 80

3.2 Re-binned cohort streamlines for the 2012 cohort. When sorted by the next grade's score, the closest 10% of students to the bin borders were moved to the neighboring bins. 92

3.3 AB cohort streamlines for the cohort of 2012. The AB process reduces regression to the mean so that the streamlines no longer converge. 94

3.4 AB snapshot streamlines for the years between 2003 and 2005. The AB process reduces regression to the mean so that the streamlines no longer converge. 96

3.5 Trajectories for the cohort of students who graduated in 2012, sorted by their 9th grade scores. The scores exhibit regression to the mean in both directions away from the binning grade. 103

4.1 AB cohort streamlines and trajectories for the cohort of 2012. With the AB process, the cohort streamlines now reproduce the trajectories quite accurately. 111

4.2 AB cohort streamlines and trajectories for the cohort of 2013. With the AB process, the cohort streamlines now reproduce the trajectories quite accurately. 112

4.3 AB snapshot streamlines of 2012-2014 and the trajectories for the cohort of students that were 6th graders in 2015, using STAAR mathematics scores. The observed data for the 2015 6th graders is considerably lower than the predicted scores using the previous cohort. 115

4.4 Comparison of score distributions for the mathematics and reading TAKS scores of the 4th graders in 2004. 117

4.5 AB snapshot streamlines of 2008-2009 sorted by reading scores and trajectories for the cohort of 2012. By using the reading score for binning, accurate predictions can be made in only two years. . . . 118

4.6 AB snapshot streamlines for 2003-2005 and trajectories for the 2012 cohort. The effects of SSI in 8th grade are captured in the trajectories but not the streamlines because the streamline data preceded SSI. 119

4.7 AB snapshot streamlines for 2007-2009 and trajectories for the 2012 cohort. Both methods capture the effects of SSI. The AB snapshot streamlines are able to predict the longitudinal data in only three years. 120

4.8 Trajectories for the cohort of 2012 disaggregated by sex. There are minimal performance disparities associated with differences in sex. 123

4.9 Trajectories for the cohort of 2012 disaggregated by SES. Students who received free or reduced lunch were considered low-income, otherwise students were considered not low-income. Not low-income students perform better than their low-income counterparts. . . . 124

4.10 Trajectories for the cohort of 2012 disaggregated by race/ethnicity. There are performance disparities that may diminish within SES groups. 133

4.11 Trajectories for the low-income students in the cohort of 2012, disaggregated by race/ethnicity. Despite having a similar SES, performance disparities still exist. 134

4.12 Trajectories for the not low-income students in the cohort of 2012, disaggregated by race/ethnicity. Despite having a similar SES, performance disparities still exist. 135

4.13 The trajectories for the cohort of 2012 disaggregated by highest level of physics course taking. Despite having similar 3rd grade scores, AP physics students outperform their basic physics and IPC counterparts. 136

4.14 Numbers of students attending AP Physics, Physics, and IPC by year. 4x4 caused a decline in IPC enrollment and an increase in basic physics enrollment. HB-5 caused a reversal of this trend. . . 137

Chapter 1

Introduction

Federal laws in the United States require primary and secondary school students in every state to take standardized assessments that act as performance and accountability measures for students, teachers, and schools. These laws have resulted in a large volume of student testing data, which is a valuable resource for education researchers who are looking to understand and influence educational policies. Education researchers have already developed many statistical methods that are used to analyze the testing data, each designed with assumptions, limitations, and strengths that allow for different types of understanding. The novel statistical method described in this dissertation aims to analyze the testing data intuitively and accurately, while directly addressing the effects of random fluctuations on the analysis.

The foundation for this new technique was developed in Marder and Bansal [7] and expounded in Bendinelli and Marder [8]. Marder and Bansal created score flow plots following from the ideas of convection and diffusion in fluid mechanics. Bendinelli and Marder continued with this idea, developing trajectories and streamlines that continuously describe the flow of student scores throughout the grade levels. By approaching the analysis of testing data

as a physics problem, Marder, Bansal, and Bendinelli contributed a different perspective that added to the statistical approaches developed in the social sciences.

While standardized test scores provide a convenient metric to study student learning, the scores incorporate random fluctuations, which can skew the results of the analysis. The random components in each score combine to create in a significant amount of regression to the mean in the streamline plots developed by Bendinelli and Marder, preventing them from accurately depicting student score flows. The alternative binning technique described in this dissertation greatly reduces the effects of regression to the mean, forming accurate score streamlines. Regression to the mean and alternative binning will be discussed in Chapter 3.

Alternatively binned (AB) streamlines can be applied to student testing data in several ways. AB streamlines can be used to study a single cohort of students longitudinally as they progress through school. In addition, AB streamlines can be constructed using several cohorts of students in a short period of time to predict longitudinal results. Lastly, AB streamlines from different periods can be compared to identify the effects of intermediate interventions.

To demonstrate the potential of this technique, AB streamlines are used to analyze Texas standardized testing data. Between 2003 and 2012, all public school students in Texas were required to take the Texas Assessments of Knowledge and Skills (TAKS). In particular, students were required to take

a mathematics exam every year from 3rd grade to 11th grade. In addition, during this period a student-level intervention, known as the Student Success Initiative, was implemented to target low-performing students. Therefore, this dataset provides a rich longitudinal setting in which the AB streamlines can be examined.

AB streamlines are being used in this dissertation within an educational context. However, this technique should be applicable in other fields with at least three repeated measurements for individuals of varying age. In particular, this technique would be useful for grouped data exhibiting regression to the mean. This technique is intended to be simple to construct and interpret so that it can be used by educators, policy makers, and researchers without statistics or physics backgrounds.

This chapter will discuss the background of standardized testing in the United States and in Texas. The dataset used in the analysis will be described, as well as the process of data collection and usage. The background of relevant education policies in Texas will be examined. Lastly, testing design and theory will be discussed. Chapter 2 will focus on some data analysis techniques already used in the social sciences and education research, as well as the techniques developed in Bendinelli and Marder [8]. Chapter 3 will describe regression to the mean, the AB process, and AB streamlines. Chapter 4 will discuss the results of the AB streamlines, as well as other results. Chapter 5 will conclude.

1.1 Standardized Testing in the United States

Standardized testing has been a common practice in the United States since the mid-19th century [9]. It has been used throughout this period to assess student performance and influence school pedagogy and curriculum. Horace Mann, one of the most influential supporters of universal public education in the 19th century, supported the idea that students should be sorted into classrooms by their tested abilities and thought that exams could be used to determine whether students were well taught [9]. A practice that was designed as a learning tool quickly became a tool to compare schools and teachers.

Technological advancements led to more efficient and reliable practices that expanded standardized testing to the masses. The first statewide high school testing program was developed in Iowa in 1929 [9]. The SAT and the ACT, developed in 1926 and 1959, still greatly influence college admission nationwide [10]. The No Child Left Behind Act of 2001 tied federal funding at the state level to student academic improvement on standardized tests [11]. Standardized tests have become a significant component of education in the United States and because their effects are far-reaching, proper analysis of the data is essential.

1.1.1 No Child Left Behind Act (NCLB)

As part of Lyndon Johnson's "War on Poverty", the Elementary and Secondary Education Act (ESEA) of 1965, particularly Title I, the largest financial component, was enacted to provide federal support for low-income

students and to create a more equitable educational environment. The declaration of the policy states [12]:

> In recognition of the special educational needs of children of low-income families and the impact that concentrations of low-income families have on the ability of local educational agencies to support adequate educational programs, the Congress hereby declares it to be the policy of the United States to provide financial assistance...to local educational agencies serving areas with concentrations of children from low-income families to expand and improve their educational programs by various means (including preschool programs) which contribute particularly to meeting the special educational needs of educationally deprived children.

The mission of providing equitable education to students in the U.S. through the ESEA has since been expanded to include not only the low-income students covered by Title I but also other underprivileged groups, such as English language learners (ELL), women, and Native Americans [11]. The ESEA has been reauthorized many times since 1965, most recently in 2001 by George W. Bush and in 2015 by Barack Obama. The ESEA has consistently been the largest federal contribution to education in the U.S. In January 2017, House Bill 610 was introduced in Congress to repeal a law that required schools to provide healthy meal options and to repeal the ESEA [13].

The No Child Left Behind Act (NCLB) was the reauthorization of the ESEA in 2001. NCLB continued to focus on the lack of educational oppor-

tunities for disadvantaged students and on closing the achievement gap. This was taken further by raising academic standards and by holding local education agencies accountable for student achievement. NCLB connected federal funding to student performance on standardized exams, both overall and for disaggregated subgroups, and imposed sanctions on students who failed to meet performance standards [11]. NCLB is known for its focus on accountability, specifically its impact on schools and teachers, and it has been the source of much controversy.

The Every Student Succeeds Act (ESSA) is the current version of the ESEA, which was implemented in 2015. The ESSA keeps the emphasis on standardized testing, although it cuts back on the federal role in establishing standards and consequences for failing to meet those standards on the exams. The power for establishing these rules is left to the states and local education agencies. The federal government can no longer require states to use the tests as part of a teacher evaluation system. However, some critics of the ESSA, including Leslie Proll, the director of policy and the NAACP Legal Defense Fund, fear that without the federal oversight, some states might not do their part in combating racial discrimination or improving education for poor children [14].

The majority of the research conducted for this dissertation involves educational data from the State of Texas between 2003 and 2012. Therefore, the laws regarding statewide assessments and their uses were set by NCLB. Under NCLB, standardized testing profoundly affected students, teachers, and

School Year(s)	2002-2004	2004-2006	2006-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014
Reading/English Language Arts	47%	53%	60%	67%	73%	80%	87%	93%	100%
Mathematics	33%	42%	50%	58%	67%	75%	83%	92%	100%

Table 1.1: Adequate Yearly Progress (AYP) standards in Texas between 2002 and 2014 [1]. This equals the percentage of students expected to reach proficiency.

schools.

1.1.1.1 School Accountability

NCLB set an ambitious goal that *every* student receiving a public primary or secondary education reach proficiency in mathematics and reading. States were expected to develop standards of proficiency and assessments to track the performance of the students. States had to establish levels of adequate yearly progress (AYP) to raise the standards of achievement on the exams until every student reached proficiency (see Table 1.1 for the AYP standards in Texas). The AYP standards are equal to the percentage of students in the state who should meet proficiency standards for the standardized exams. Under NCLB, states were expected to bring the proficiency of students in mathematics and reading up to 100% by 2014 [1]. The AYP was not only judged based on standardized test performance, but also on test participation and graduation rates.

In addition to setting standards of achievement, NCLB set consequences for schools that failed to show AYP [15]. If a school failed to meet AYP standards for two consecutive years, they were Identified for Improvement. This

required the school to notify the parents of students, allowing the parents to choose to send their children to a different school. The school also needed to develop an improvement plan and to designate 10% of their Title I funds for professional development. If the school failed to meet AYP standards for four years, the school required Corrective Action. They were required to implement at least one of the following options:

- Implement research-based curriculum or instructional program
- Decrease school's management authority
- Extend school day or school year
- Restructure school's organization
- Replace staff relevant to school's low performance
- Appoint an outside expert

If the school did not show AYP for a fifth year then the school was restructured by one of the following interventions:

- Reopen school as a charter school
- Replace all or most staff
- Contract with another entity
- Yield to state takeover of school

- Conduct other major governance restructuring

A school would have to meet AYP for two consecutive years to correct its status [15].

The severity of these standards and consequences is evident by the number of schools that failed to meet AYP standards. Of the 8,529 campuses in Texas in 2012, only 40% met the AYP standards, while 48% failed to meet the standards (some were not evaluated) [16]. Almost all of those schools failed to meet the AYP standards because of poor performance on the standardized exams. On September 30, 2013, Texas was granted a waiver which exempts the state from AYP requirements. By March 2014, 42 states had waivers [17]. In 2015, Texas was given a high-risk status for failing to meet ESEA standards, even with the added flexibility of the waiver [18].

1.1.1.2 Value-Added Modeling

School teachers were greatly affected by NCLB and the accountability measures. Most directly, NCLB required that teachers be "highly qualified". This meant that the teachers needed to have a bachelor's degree, they needed to be fully certified by the state, and they needed to demonstrate competency in each core academic subject that they taught. The U.S. Department of Education claimed that the impetus for higher teacher qualification standards was a series of studies conducted in Tennessee and Texas that found that "the students who had effective teachers greatly outperformed those who had ineffective teachers" [19]. These studies used value-added modeling to determine

the impact of the teachers on the students' scores.

Value-added modeling (VAM) is a technique that is used to isolate the contribution of individual teachers to the changes in score of their students. A common form for a VAM [20] is:

$$Y_{isjt} = \beta_0 + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + S_{isjt}\beta_3 + T_{isjt}\theta + \epsilon_{isjt} \tag{1.1}$$

where Y_{isjt} is the score for student i at school s with teacher j in the year t, X_{isjt} is a vector of student characteristics, S_{isjt} is a vector of school/classroom characteristics, T_{isjt} is a vector of teacher characteristics, and ϵ_{isjt} is a random term. The teacher value-added parameters are contained in θ. VAMs are longitudinal mixed-models, which will be discussed further in chapter 2.

There are several criticisms of VAM, particularly the use of VAM in teacher evaluations. VAM can be biased because students are not necessarily randomly assigned to teachers and non-random sorting processes are not always documented or observed so that they could be controlled in the model [21]. VAM can also be unstable in the sense that the results do not always correlate well for the same teacher across years or classrooms [20]. In addition, they do not consider latent contributions such as parental involvement.

Despite this criticism, VAM consistently measures substantial variation in teacher performance [20]. As a result, VAM is often used to inform personnel decisions. There are several studies that have measured the benefits of raising the quality of teachers in schools based on value-added scores. This

is done either by replacing low-performing teachers [22, 23, 24] or reassigning teachers to other schools/classrooms based on their performance [25, 26]. Furthermore, some teacher VAM scores have been shared publicly to influence school choice decisions [27]. The use of value-added modeling to evaluate teacher performance and manipulate teacher assignments is an example of the impact of standardized exams on teachers and the importance of accurate and reliable score analysis.

1.2 Texas Setting

1.2.1 Education Research Center (ERC)

Although the methods of this dissertation are expected to have general utility, the possibility of testing them is due to a unique resource provided by the Texas Education Research Center (ERC), which maintains extensive longitudinal information about Texas students, teachers, and schools. The ERC is officially designated by the State of Texas as a research center for the purposes of policymaking and scientific inquiry [28]. According to the ERC's introductory webpage,

> Since its inception in 2006, the Texas ERC's goal has been to bridge the gap between theory and policy by providing a cooperative research environment for study by both scholars and policy makers. As part of its mission, the Texas ERC works with researchers, practitioners, state and federal agencies, and other policymakers to help inform upon critical issues relating to education today. [28]

The ERC dataset includes all the public school information from pre-kindergarten to 12th grade collected by the Texas Education Agency (TEA), as well as information from beyond high school collected by the Texas Higher Education Coordinating Board (THECB) and the Texas Workforce Commission. In addition, there are some data collected on a national level, including some from the National Student Clearinghouse and the National Center for Educational Statistics. It is one of the largest State Longitudinal Data Systems (SLDS) in the country [28].

The TEA provides the ERC with the bulk of the data pertaining to the schools and students in Texas that are used in this dissertation. On a campus and district level, financial and staff employment data are provided. On the student level, enrollment, course taking, special educational needs, and demographic data are provided. Additionally, the scores from statewide assessments are provided, including the Texas Assessment of Academic Skills (TAAS), Texas Assessment of Knowledge and Skills (TAKS) and the State of Texas Assessment of Academic Readiness (STAAR).

The State Board of Educator Certification (SBEC) data, also included in the ERC dataset, includes information about all of the teachers in the State of Texas and the details of their teaching certification. Records include information about where they were certified, the field and type of certification, and the time frames for the certifications. This data, combined with the staff employment records and student course completion records from the TEA, allows for a connection between student outcomes and teacher preparation.

This connection is the focus of the research proposal that prompted our access to the ERC data.

1.2.1.1 ERC Access: FERPA

While there is ample educational data that is made publicly available, access to the ERC's database requires approval from the state employees at the ERC. The main reason for the limited access is the Family Educational Rights and Privacy Act (FERPA) of 1974. FERPA is a federal law that protects the privacy of student educational records collected by schools that receive funding from the U.S. Department of Education [29]. FERPA grants access of underage (< 18) student education records to parents and approved officials. FERPA also limits the data schools can make publicly available to just "directory" information, including name, address, phone number, birth date, awards, and attendance. Parents can request that this information be kept private.

To comply with FERPA regulations, there are several tactics utilized by the ERC that limit the use and distribution of the data. The data is *de-identified*, meaning the student and teacher names are removed from the dataset and replaced with identification numbers, without a connection between the numbers and the names. Still, the data could be presented in such a way that the identities of the students could be recovered through quasi-identifiers. Therefore, the ERC must approve all outgoing information before it can be disseminated. This information must be *masked* so that data are not

reported for any group of students containing less than five members.

The work required to analyze the data before it is ready for ERC approval and release must be completed in a protected environment. Researchers are provided access to locked offices containing computers that remotely access the ERC server. Researchers cannot add or remove anything from these computers and there is no Internet access. Any necessary software must be added to these computers by the ERC Information Technology professionals. The computers and offices are shared amongst all the approved researchers, so computer time must be scheduled.

Before access can be gained to the ERC database, researchers must submit a proposal to be approved by the advisory board. The proposal must meet a "minimum standard of rigor and [provide] a benefit to the education in the state" [30]. Furthermore, the approval of proposals is often influenced by political agendas. Researchers must submit proof of masking and FERPA training, Institutional Review Board (IRB) approval, and a confidentiality agreement. If the proposal is approved, the relevant data files (determined by the often cryptic variable names provided online) must be requested and their use justified. Therefore, researchers are only provided access to a subset of the available data. Additionally, access to the ERC database requires an annual payment [28], which is used to fund the maintenance of the server and database, and to support the employees who review the material submitted for release.

Our proposal with the ERC is to study the student outcomes of gradu-

ates of UTeach and other teacher preparation programs. UTeach is a secondary STEM teacher preparation program that was started at the University of Texas at Austin in 1997 and has since expanded to 45 other universities [31]. The Principal Investigator for the research project is Michael Marder. The original proposal was submitted in 2014. Since then, several researchers joined the project including Caitlin Hamrock, a former sociology doctoral student, Matthew Guthrie, a physics doctoral student, Bernard David, a STEM education doctoral student, and myself. Dhruv Bansal, Anthony Bendinelli, and David McGhan are former physics doctoral students who were involved in the ongoing research questions but were not on the ERC proposal. Guthrie and I were added to the proposal in the summer of 2015. Our additional contribution to the project included longitudinal data analysis methods, school comparison investigations, and questions involving equity in education, particularly in the fields of Science, Technology, Engineering, and Mathematics (STEM). In the process of investigating educational disparities using the longitudinal methods described below, I developed a new method that improved upon the inherited methods. Therefore, while other researchers on our proposal were focused more directly on addressing the questions related to teacher preparation programs and their effect on student outcomes, the research in this dissertation focuses more on methodology and the analysis of standardized exam scores. Recently, two additional proposals with the ERC were approved; one studies the impacts of Texas House Bill 5 and one focuses on the outcomes of the Rio Grande Valley Linking Economic and Academic Development (RGV LEAD)

project. My work on these projects is ongoing.

1.2.2 Texas Standardized Exams

Since 1979, when assessment programs were first implemented in Texas, there have been five standardized tests that have been used to assess the academic performance of Texas students [32]. In 1980, the Texas Assessment of Basic Skills (TABS) was implemented as a result of a law enacted by the 66th Texas Legislature, which required students to demonstrate basic skills in mathematics, reading, and writing. In 1986, the Texas Educational Assessment of Minimum Skills (TEAMS) was implemented by the Texas Education Agency (TEA) and was the first exam in Texas that required a passing grade to be eligible for high school graduation. In 1990, the Texas Assessment of Academic Skills (TAAS) was first implemented, expanding the testing to more grades and subjects. In 2003, the Texas Assessment of Knowledge and Skills (TAKS) replaced TAAS as the primary testing program. TAKS was designed to measure students' understanding of the statewide curriculum, known as the Texas Essential Knowledge and Skills (TEKS). In 2012, the State of Texas Assessments of Academic Readiness (STAAR) was introduced as the fifth and current standardized test in Texas. This change mostly affected high schoolers as it replaced the grade-specific exams after 8th grade with end-of-course (EOC) exams.

The oldest data available to us at the ERC dates back to the 2002-2003 school year, therefore the standardized exams that will be focused on in

Grade	Math	Reading	English Language Arts	Writing	Social Studies	Science
Grade 3	✓	✓				
Grade 4	✓	✓		✓		
Grade 5	✓	✓				✓
Grade 6	✓	✓				
Grade 7	✓	✓		✓		
Grade 8	✓	✓			✓	✓
Grade 9	✓	✓				
Grade 10	✓		✓		✓	✓
Exit Level	✓		✓		✓	✓

Table 1.2: TAKS exam subjects by grade [2].

this dissertation are TAKS and STAAR. In particular, the nine year period between 2003 and 2012 is a good setting for research because the TAKS exam was fairly consistent in this period and there were several interventions during this period with outcomes that can be studied using the TAKS scores.

TAKS encompassed exams from many subjects for public school students starting in grade 3 through high school. The students were tested in mathematics in grades 3-10 and exit level; reading in grades 3-9; writing in grades 4 and 7; English language arts in grades 10 and exit level; social studies in grades 8, 10, and exit level; and science in grades 5, 8, 10, and exit level [2]. Table 1.2 provides a summary of the subjects assessed in each grade. Students were therefore required to take up to four subject exams in a school year, specifically between March and May (with the exception of retesting in June/July).

The exit level exams were a large component of the TAKS program.

Exit level exams were usually taken in 11th grade, but could be taken in other grades. Students needed to pass the four exit level subject exams—mathematics, English language arts (ELA), science, social studies—to be eligible for a public high school diploma. Texas law required that the material on the four exams include Algebra I and Geometry, English III and Writing, Biology and Integrated Physics and Chemistry (IPC), and Early American and U.S. History, respectively [2]. Senate Bill 103 also mandated that the TEA incorporate a college readiness component into the exit level TAKS exams. The TEA, its testing contractor Pearson Educational Measurement, and the Texas Higher Education Coordinating Board (THECB) collaborated to create a Higher Education Readiness Component (HERC) score on the exit level exams. If a student had a passing HERC score on the exit level TAKS, then they were exempt from having to take the Texas Success Initiative Assessment (TSIA), which is a requirement for college course enrollment in Texas [33]. Military and transfer students are also exempt, as well as students who meet Texas Success Initiative (TSI) standards on the ACT or SAT. Students who are not exempt from the TSIA must take the exam if they want to enroll in college courses. If students fail the exam, they must take remedial coursework, which can cost the same amount as a college course in tuition but does not count toward a degree.

Accommodations for the TAKS exams were provided to students who had special needs as mandated by NCLB [2, 32]. In 2001 until 2007, State-Developed Alternative Assessments (SDAA) were used as an alternative as-

sessment for students receiving special education services. TAKS-Inclusive was implemented between 2006-2008 and it covered the subjects tested by TAKS that were not in the SDAA. In 2008, the testing accommodations were reorganized using several testing alternatives. TAKS (Accommodated) was provided to eligible students who were held to the same learning expectations but required adjustments to the exam in its presentation, setting, timing, or the format of the students' responses. TAKS-Modified was a grade-level exam taken by students receiving special education services and it had adjustments in the format and test design. Eligible students did not have to pass the TAKS-Modified to graduate high school. TAKS-Alternate was an exam for students with significant cognitive disabilities and it involved teacher observation rather than a traditional multiple-choice test. All of these students, except those taking TAKS-Alternate, were subject to the grade promotion requirements mandated by the Student Success Initiative, described below.

NCLB also mandated that accommodations be made for English language learners (ELLs). Spanish versions of TAKS were available for ELLs or students with limited English proficiency (LEP) for grades 3-6 until 2009, and thereafter only grades 3-5. Linguistically accommodated testing (LAT) was available for eligible ELL immigrants in all grades except for the exit level exams. The implementation of LAT was staggered with mathematics starting in 2005, reading/ELA in 2007, and science in 2008 [32]. The Spanish versions were designed to align with the English versions, holding the students to the same academic standards.

In 2007, Senate Bill 1031 aimed to replace high school grade-specific exams with end-of-course (EOC) exams, phasing out some TAKS exams beginning with the 9th graders in the 2011-2012 school year. In 2009, House Bill 3 passed, which called for the complete replacement of TAKS with STAAR, beginning in Spring 2012. For students in grades 3-8 the tested subjects and grades remained the same. High school exams were replaced with 15 EOC STAAR exams in Algebra I, Geometry, Algebra II, English I reading and writing, English II reading and writing, English III reading and writing, Biology, Chemistry, Physics, World Geography, World History, and U.S. History. Students were required to meet passing standards on 11 of the 15 exams, with a minimum cumulative score in each content area required for graduation. If a student met TSI standards on the Algebra II and English III exams, then they were declared *college ready* and were exempt from having to take the TSIA. House Bill 5 (HB-5), which passed in 2013 and was implemented in 2014, reduced the number of EOC exams from 15 to 5 keeping Algebra I, English I (combined reading and writing), English II (combined reading and writing), Biology, and U.S. History.

The reduction in EOC exams due to HB-5 had profound implications for the college readiness of Texas students. The Texas Education Code still lists the TSIA exemption STAAR requirements in terms of the Algebra II and English III EOC exam scores [34]. Students are not required to take these exams for graduation and therefore the exams are not widely available. When TAKS was the primary standardized exam, students were already required to

take the exit exams for graduation, so those who performed at the TSI standard were automatically considered college ready. There was no extra effort on the part of the students. Now students have to take extra courses and exams to become exempt from the TSIA through the standardized tests. The consequences of House Bill 5 on college readiness can be seen in Figure 1.1. College readiness in Texas was increasing at a steady rate between 2006 and 2013 for schools in each poverty quartile. In 2014, when HB-5 was first implemented, the direction of growth begins to change with a large decrease in college readiness by 2015, most dramatically affecting the more impoverished students.

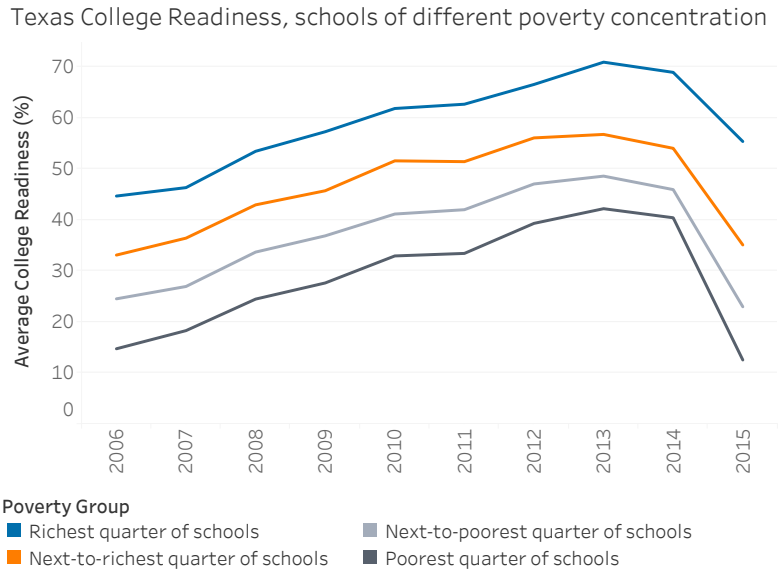

Figure 1.1: College readiness in Texas schools by poverty quartile. The drop in college readiness by 2015 is likely due to the exam requirement changes since HB-5. Figure provided by Marder.

School Year(s)	Program
1999-2000 to 2002-03	Teacher Reading Academies (K-3)
1999-2000 to 2008-09	Accelerated Reading Instruction
2000-01 to 2001-02	Teacher Math Academies (5-7)
2003-04 to 2008-09	Accelerated Math Instruction
2003-04 to 2008-09	Intensive Reading Instruction
2005-06 to 2008-09	Intensive Mathematics Instruction
2007-08 to present	Texas Adolescent Literacy Academies
2009-10 to present	Student Success Initiative Grants
2009-10 to present	Rider 42 Professional Development Academies
2009-10 to present	Algebra Readiness Grant

Table 1.3: Timeline of Student Success Initiative (SSI) programs by school year [3].

1.2.3 Student Success Initiative (SSI)

Since its enactment in 1999 by the 76th Texas Legislature, the Student Success Initiative (SSI)—an umbrella program which includes several smaller initiatives. (see Table 1.3)—aims to help all Texas students perform at grade-level in mathematics and reading. The encompassed initiatives belong in several categories; some establish standards for satisfactory performance, some are interventions for students who fail to meet those standards, and others provide professional development for teachers so that unsatisfactory performance may be prevented. The implementation of these strategies has varied over the years and so has the budget, to a great extent (see Table 1.4). The grades impacted by the SSI grew along with the cohort of students graduating high school in 2012.

One major component of the Student Success Initiative is a set of grade promotion requirements that mandate especially high-stakes standardized ex-

School Year	Funding Level	Grades Impacted
1999-2000	$65.99 million	Kindergarten
2000-01	$107.29 million	Kindergarten-Grade 1
2001-02	$110.28 million	Kindergarten-Grade 2
2002-03	$120 million	Kindergarten-Grade 3
2003-04	$82.35 million	Kindergarten-Grade 4
2004-05	$82.35 million	Kindergarten-Grade 5
2005-06	$158.01 million	Kindergarten-Grade 6
2006-07	$158.01 million	Kindergarten-Grade 7
2007-08	$154.50 million	Kindergarten-Grade 8
2008-09	$154.50 million	Kindergarten-Grade 8
2009-10	$152 million	Kindergarten-Grade 12
2010-11	$152 million	Kindergarten-Grade 12
2011-12	$20.5 million	Kindergarten-Grade 12
2012-13	$20.5 million	Kindergarten-Grade 12
2013-14	$25.25 million	Kindergarten-Grade 12
2014-15	$25.25 million	Kindergarten-Grade 12
2015-16	$15.85 million	Kindergarten-Grade 12
2016-17	$15.85 million	Kindergarten-Grade 12

Table 1.4: Total appropriated funding and impacted grades for the SSI by school year [3, 4]. The budget has decreased substantially in recent years.

ams in 5th and 8th grade. Students can have up to three attempts to pass their reading and mathematics exams in 5th and 8th grade. If they fail all three tries in either subject, the student is retained in that grade for another year unless they are cleared by a committee [35]. There was also the possibility for grade retention due to the 3rd grade reading exam but that was eliminated by House Bill 3 in 2009. The grade promotion requirements were first implemented for the cohort of students who graduated high school in 2012 (see Table 1.4 for the SSI timeline). More precisely, the first 3rd grade requirement was for the 3rd graders in the 2002-2003 school year, the first 5th grade requirement was for the 5th graders in the 2004-2005 school year, and the first 8th grade requirement was for the 8th graders in the 2007-2008 school year [3].

There have been several methods used under the umbrella of SSI to combat unsatisfactory scores on the standardized exams. One of the methods was to identify struggling students who were at risk of failing the exams and then provide additional instruction to these students. The main programs that provided this targeted intervention at the early stages of the SSI were the Accelerated Mathematics Instruction (AMI) and the Accelerated Reading Instruction (ARI). Students were selected for these programs if they had already failed the first or second administration of the exams, or if they performed poorly on other diagnostic tests. ARI was first implemented for just kindergarten in the 1999-2000 school year. Each subsequent year, the program was expanded to include an additional grade, so that by 2008, ARI was provided

to struggling reading students in K-8th grade. In the 2003-2004 school year, AMI was first implemented for K-4th grade and was also expanded each year so that it was helping K-8th grade mathematics students by 2008 [36].

In the 2006-07 school year for grades K-7 (those impacted by SSI), 29% of the students were struggling in reading and 25% were struggling in mathematics. Of those students, 79% of the reading students received ARI funding and 82% of the mathematics students received AMI funding. Of *those* students, 69% of the reading students were performing at grade-level after ARI and 68% of the mathematics students were performing at grade-level after AMI [36]. Both the AMI and the ARI were dismantled in 2009. Accelerated instruction is still a component of SSI, although the type of instruction is decided at a local level. This additional instruction may take place during normal school hours, after school, or during the summer [35].

We have applied the new methods described in this dissertation to the mathematics TAKS scores between 2003 and 2012. In this period, SSI was implemented in stages along with the cohort of 2012. Therefore, we use the methods to study the combined effects of SSI and AMI/ARI. More current results could be studied using STAAR scores, although the transition from TAKS to STAAR complicates the analysis. The proposed 2018-2019 biennium budgets by the Texas House and Senate cut funding to education considerably, and in particular, the Senate's initial proposal cut SSI entirely [37]. The most recent version of the budget, Senate Bill 1, lists a $5.5 million annual budget for SSI [38] and took effect in September 2017.

1.3 Testing Theory and Design

Standardized tests exist to gauge the knowledge and skills of the test takers. It is therefore necessary for the scores on the exams to at least partially characterize this underlying ability. However, testing performance does not accurately or fully represent knowledge and skills.

There are many sources for the discrepancy between the test scores and the knowledge it is testing. The finite number of test items on an exam means that it cannot assess the breadth of a student's knowledge. On multiple choice exams especially, students do not need to know the correct answer to answer correctly. A student without any indication of the correct answer would have a 25% chance of guessing correctly if there were four answers to choose from. Some knowledge would allow the student to narrow down the options, increasing their chances of answering correctly. Students may make mistakes when submitting their answers, or the stressful conditions of standardized exams may prevent them from utilizing their knowledge to the full extent. There could be issues when scoring the exams, or questions may be worded poorly. The tests could be biased, benefiting some students over others regardless of knowledge.

These sources of discrepancy, be it luck, test design, or something else, contribute an error component in the test scores. Short-term random fluctuations, in particular, are of great concern when interpreting the exam results. Statisticians and psychometricians have developed many methods for dealing with these random fluctuations. The technique described in this dissertation

aims to address regression to the mean due to these random fluctuations in standardized test scores.

1.3.1 Classical Test Theory

Classical Test Theory, or True Score Theory, is a simple and commonly used theory that relates the measured test score and the random error to a student's true score. Classical Test Theory was codified in its current form in the 1960s by Melvin Novick and Frederic Lord [39, 40].

The central idea of Classical Test Theory is that test scores are flawed measurements of testing performance. Given a student's knowledge and test taking ability, that student is expected to get a particular score on an exam, which is called their *true score*. More precisely, if a student were to take infinite administrations of a test, the expectation value of those scores would be their true score. On a given administration, due to scoring error, luck, or other factors, their observed score would vary from this true score. Given an observed score X, a true score T and an error score E, the relationship in Classical Test Theory is given by:

$$X = T + E \tag{1.2}$$

For an individual that is tested repeatedly with *parallel measurements* (measurements that have equivalent true scores and observed score variances), the true score would be a constant value while X and E would be random variables. The error scores are assumed to be uncorrelated with each other

and with the true scores. The expectation value of the observed score is equal to the true score and the expectation value of the error score is zero. These assumptions lead to a simple relationship between the variances of these variables [40].

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \tag{1.3}$$

This equation implies that the variance in the observed scores comes from two sources: variance in the true scores and variance due to the random fluctuations. This leads into the concept of *reliability*–how does the variation in observed scores compare to the variation in true scores. The reliability ϱ is defined as the ratio of the true score variance to the observed score variance and is equal to the square of the correlation between the true and observed scores [40].

$$\varrho = \rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2} \tag{1.4}$$

While a useful definition, since the true scores and variances are unmeasured, the reliability cannot be calculated with a single exam. Parallel measurements can lead to a measurable metric of reliability. The correlation between the observed scores from two parallel exams is in fact equal to the reliability [40] and so,

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{1.5}$$

Therefore, given the correlation between two parallel exams and the variance in the observed scores, one can learn a considerable amount about the exam. Let's say, as an example, that the correlation between parallel exams equaled 0.81 and the (observed) standard deviation on the exam was 10 points. The

standard deviation in the true scores would be 9 points, 81% of the variance in the observed scores would be due to variance in the true scores, and the correlation between the true and observed scores would be 0.9.

Of course, parallel exams are more of an idea than a reality. However, by utilizing the composite nature of the exams, the fact that there are several test items, the lower bound for the reliability can be established. This lower bound is called the Cronbach coefficient α and is defined as:

$$\alpha \equiv \frac{N}{N-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right] \tag{1.6}$$

where N is the number of items, σ_i^2 is the variance of item i, and σ_X^2 is the variance of the observed scores [41]. A special case of the Cronbach α is used to compute the lower bound for reliability on the TAKS exams, as discussed further below.

1.3.2 Item Response Theory

Item Response Theory takes the concepts in Classical Test Theory a step further. Item Response Theory is less concerned with the aggregate test score and instead focuses on the responses on the individual questions or items on the exam. For a given question, the probability of answering correctly would depend on the latent abilities of the students, and this probability distribution may differ for each item.

A student with latent ability θ would have some probability $P(\theta)$ of answering an item correctly. For that item, students with varied abilities

will have varied probabilities and this relationship between the ability and the probability of answering correctly for a given item is called the *item characteristic curve*. Often this relationship is modeled as an S-shaped curve, although it could take any shape, including a step function. Therefore, increasing ability usually results in a higher probability of answering correctly, although for the most and least able students, small changes in ability will not change the probability as much as changes in ability for students in the middle of the distribution.

The item characteristic curve gives a lot of information about the item including its difficulty and its discriminatory abilities. The location of the higher sloped section of the S curve determines the difficulty of the item; items that have quickly increasing probabilities at lower abilities are easier and items that have increasing probability at higher abilities are harder. The magnitude of the slope determines the discriminatory ability of the item; items with a steep slope are more discriminatory as they effectively set a threshold in ability for correct answers whereas items with a small slope are less discriminatory and students within a large range of abilities have similar probabilities of answering correctly.

To describe the difficulty and discrimination of the item characteristic curves more precisely and rigorously, several models are used in item response theory to describe the curve. The two-parameter logistic model is one of the preferred models, as it is simple and it has parameters for the difficulty and discrimination. The probability of answering an item correctly as a function

of the student ability θ is

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}} \ , \tag{1.7}$$

where a is the discrimination parameter[1] and b is the difficulty parameter [43].

A simplified version of the two-parameter logistic model is the one-parameter model, where the discrimination parameter is the same for every item. This is also known as the Rasch model. The probability function becomes

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}} = \frac{e^{\theta - b}}{1 + e^{\theta - b}} \ , \tag{1.8}$$

where θ is a measure of ability and b is the difficulty of the item (note that the parameters are not necessarily equivalent to those in the two-parameter model even though the same symbols are used). The Rasch model is the model that TEA uses to scale the exams, as discussed further below.

Another of the commonly used models in Item Response Theory adds a third parameter to compensate for any guessing that takes place on a multiple choice exam. The three-parameter logistic model gives the probability of answering an item correctly in terms of the student ability θ as

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \ , \tag{1.9}$$

where a and b are again the discrimination and difficulty parameters and c is the guessing parameter [43]. The parameter c is the probability of getting the

[1] In much of the item response theory literature, the logistic curve is scaled by 1.702 to make the logistic ogive similar to a normal ogive and the exponential term is written $e^{-da(\theta - b)}$ where d is the scaling constant [42].

correct answer purely by guessing and it is not a function of the ability. For the one and two parameter models, the value of b corresponded with the value of θ where the probability of answering correctly was 0.5. In the three-parameter model, the parameter c gives a minimum probability of answering correctly (typically 0.25) and therefore the difficulty parameter b instead equals the ability value where the probability is halfway between c and 1. For the two-parameter model, the slope of the item characteristic curve when the ability equals b (and the probability is 0.5) is equal to $a/4$. For the three-parameter model, the slope of the item characteristic curve when the ability equals b is equal to $a(1-c)/4$.

The TEA uses the Rasch model (one-parameter logistic function) to express the probabilities of answering the TAKS items correctly. Under the Rasch model, several expressions can be defined to learn information about the exam and the abilities of the test takers. Given the probability function in Equation 1.8, a student's true score as a function of their ability can be defined as

$$T(\theta) = \sum_{i}^{N} P_i(\theta) \ , \tag{1.10}$$

which is simply the sum over the N items in the exam of the probability of answering the item i correctly [43]. The true score function is also called the test characteristic curve, since it represents the most probable scores for students of varying ability. For the Rasch model (and the two-parameter logistic model) a null score corresponds with a negatively infinite ability and a perfect score corresponds with an infinite ability. Essentially, the test characteristic

curve transforms the ability score into a true score, which can be interpreted more easily since it predicts the number of questions answered correctly.

1.3.3 TAKS Test Design

During the design and evaluation of TAKS and STAAR, the TEA follows the guidelines set by the Standards for Educational and Psychological Testing, which was developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The TEA uses standard setting, scaling, and reliability measures to ensure that the exams are reliable, fair, and accurate.

For both TAKS and STAAR, the TEA uses a twenty step process for test development [32]. Educators and administrators collaborate with the TEA to create a list of grade and subject specific categories inspired by the state-mandated curriculum to be tested. Test items are developed from these categories by the TEA and Pearson Educational Measurement. Informed by field-tests, these test items are reviewed and revised to create a bank of items with similar content and difficulty. The exams are built using this bank of test items such that the exams are equivalent in difficulty across administrations.

The TEA sets performance standards to quickly determine how well students meet the expectations set by the TEKS curriculum. With TAKS, there were two cut scores that split the students into three categories: commended performance, met standards, did not meet standards. The number of correct answers that corresponded to these cut scores varied with each test ad-

ministration. The TEA computed scaled scores for each exam to adjust for the variation between each administration. They have used both horizontal scaling and vertical scaling. Horizontal scaling compares exams within a grade and subject, finding equivalent scores from different versions of the exam over time. Vertical scaling compares exams across grade levels within the same subject, allowing for the development of growth measures for individual students. The TEA switched from horizontal scaling to vertical scaling in spring 2009 [44], approximately halfway through the period studied in this dissertation. Scaled scores will be discussed further below.

In theory, a reliable exam would produce consistent scores for a population of students if they were retested multiple times. In practice, students cannot be retested to determine the reliability so internal consistency measures from a single test administration are used instead. For multiple choice exams, such as the mathematics TAKS exams, the TEA uses the Kuder-Richardson 20 (KR20) formula, a special case of the Cronbach α when the components are dichotomous (right or wrong), to calculate the lower-bound estimate of the true reliability [2]:

$$\alpha_{(20)} = \left[\frac{N}{N-1}\right]\left[1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{\sigma_X^2}\right], \tag{1.11}$$

where N is the number of items on the test, σ_X^2 is the observed score variance on the test, and p_i is the proportion of students who answered item i correctly.

Grade	Mathematics KR20 α
Grade 3	0.878
Grade 4	0.888
Grade 5	0.902
Grade 6	0.909
Grade 7	0.904
Grade 8	0.907
Grade 9	0.924
Grade 10	0.918
Grade 11	0.904

Table 1.5: Lower bounds for the reliability on the TAKS mathematics exams in 2010 [2].

Table 1.5 shows the KR20 values for the mathematics TAKS exams in 2010 as calculated by the TEA.

The standard error of measurement can be calculated using the KR20 estimate of reliability and the standard deviation of the observed scores:

$$\text{SEM} = \sigma\sqrt{1 - \alpha_{(20)}}. \tag{1.12}$$

The tests are constructed using the Rasch model discussed in the previous section (Equation 1.8). Since the difficulty parameters and students' abilities are unknown, the tests must be calibrated. With the Rasch model, students with the same raw score will have the same estimated ability and test items with the same number of correct responses have the same difficulty. The proof is achieved by comparing the log-odds for two students on the same item or two items for the same student. The logit (log-odds) of the probability

function for a student with ability θ on an item with difficulty b is

$$\text{logit}(P(\theta)) = \log\left(\frac{P(\theta)}{1-P(\theta)}\right) = \theta - b. \tag{1.13}$$

The difference in the log-odds for students with different abilities on the same item is independent of the item's difficulty and only depends on their abilities. Similarly, the difference in log-odds for a student on two items with different difficulties would be independent of the student's ability and only depend on the difficulties of the items. Therefore, students with the same raw score will have the same estimated ability and test items with the same number of correct responses have the same difficulty.

The process of calibration is complicated and involves many iterations to alternatively adjust the item difficulty and ability parameters [45]. For a single student, given the best estimation for the difficulty parameters, the following equation could be used iteratively to estimate their ability:

$$\hat{\theta}_{s+1} = \hat{\theta}_s + \frac{\sum_i^N u_i - P_i(\hat{\theta}_s)}{\sum_i^N P_i(\hat{\theta}_s)(1-P_i(\hat{\theta}_s))}, \tag{1.14}$$

where $\hat{\theta}_s$ is the estimated ability during iteration s and u_i is a binary variable equaling one when the answer for item i is correct and zero when the answer is incorrect [43, 45]. This is repeated until the adjustment term is minimized.

The standard error of the estimated ability is calculated as [43]

$$SE(\hat{\theta}) = \left(\sum_i^N P_i(\hat{\theta})(1-P_i(\hat{\theta}))\right)^{-1/2} = \left(\sum_i^N I_i(\hat{\theta})\right)^{-1/2}, \tag{1.15}$$

where $I_i(\theta)$ is the item information function. This process does not work for estimating the ability of students with perfect or null scores. Once this process has been performed iteratively until the values are optimized, the estimated abilities are used to calculate the scaled scores, which can be mapped to the corresponding raw scores. The expressions used to calculate the scaled scores from the abilities are discussed below.

1.3.4 Scaled vs. Raw Scores

Before the spring of 2009, TAKS was scaled horizontally, meaning the exams were comparable within a grade and subject but not between grades. The benefit of a horizontal scaling is that the scaled cut scores that determine the scores labeled Met Standard level (2100) or Commended Performance level (2400) are the same each year. However, longitudinal comparisons cannot be done using horizontally scaled scores. The linear transformation the TEA uses to calculate the horizontally scaled scores from the ability scores determined by the Rasch model is

$$S_j = T1\theta_j + T2, \tag{1.16}$$

where S_j is the scaled score for student j given an ability θ_j, and $T1$ and $T2$ are constants provided by TEA that are unique for each grade, subject, and year of the exams.

Starting in the spring of 2009, as a result of changes to the Texas Education Code, the TEA switched from a horizontal scaling method to a vertical scaling method [44, 2] for the reading and mathematics exams in grades 3-8.

This would allow for scores in the same grade and subject to be compared from year to year. The cut scores would change each year. The linear transformation used to calculate the vertically scaled scores from the Rasch model abilities is

$$S_j = T1\theta_j + T1LC + T2, \tag{1.17}$$

where S_j is the scaled score for student j given an ability θ_j, $T1$ and $T2$ are constants that are the same for every grade and year within a subject, and LC is a constant that is the same every year within a subject and grade. The constants are provided by the TEA without details about their generation.

In our research, we have decided to use the raw percent score instead of the scaled scores provided by the TEA. While psychometrically, scaled scores are preferable to raw scores when comparing the results from multiple exams, we feel justified in using raw scores for several reasons. First, the method that TEA uses to compute the scaling constants is proprietary and therefore the scaled scores are less transparent than the raw scores. Second, the iterative process of estimating the abilities using the Rasch model requires a computer with software designed for item response theory calculations. Therefore the scaled scores are less understandable and intuitive than the raw scores. Third, the use of the Rasch model means that raw scores, abilities, and scaled scores are all mapped one-to-one-to-one so the varied difficulty of the items does not affect the scoring methods differently. Fourth, the scaling was never vertical in every grade and the exams in grades 3-8 switched scaling methods in the middle of the relevant period. On the other hand, the total possible score for

the exams within a grade and subject was consistent each year. Therefore the scaled scores were less consistent than the raw scores and could not be used for longitudinal research. Fifth, the exams were made using a bank of items that were tested and designed to be similar in content and difficulty. Each year the exams were constructed to be very comparable. Therefore the raw scores were designed to be fairly consistent metrics from year to year. Sixth, the raw score to scale score conversion tables were examined and the raw score corresponding with a passing score within each grade for the mathematics TAKS exams was fairly consistent; in grades 3-8, the score varied by at most one question, and in grades 9-11 the score varied by at most two questions (except for the slightly more difficult 9th grade exams in 2010 and 2011, which varied by three questions). Therefore, on the basis of transparency, simplicity, and consistency, we have chosen to use raw score percentages as our metric.

Chapter 2

Background

Longitudinal data analysis methods have been developed in many fields to study both the within-person and between-person changes over time [46]. Within-person changes are particularly pertinent to the analysis of educational policy impacts because they can establish a connection between an intervention and the resulting changes for the affected individuals. Between-person changes are also important to study because they can show how interventions affect students on a larger scale, and they can shed light on the differential impacts of an intervention on different groups of students, which can lead to more efficiently targeted or more equitable interventions.

Most longitudinal studies utilize a statistical technique known as multilevel modeling (also called hierarchical linear modeling) [47, 48], or similar techniques in structural equation modeling [49, 50, 51], to analytically describe both the within-person change and the between-person change in the longitudinal data. The analytical model may use observed variables or unobserved latent variables combined with fitted parameters to best represent the growth of the outcome variable. Individual growth curves or latent growth curves represent the within-person component of the model, and person-specific pa-

rameters (usually within a normal distribution of values) that adjust the intercepts or slopes account for the between-person variation. The model most often takes the form of a linear, polynomial, or piece-wise function, although it can have any non-parametric form.

This idea can be extended to grouped individuals through techniques called group-based trajectory modeling [52] and latent class growth modeling [53]. These techniques assume that the population is mixed, containing several distinct groups that are categorized by an unknown variable. Using the longitudinal data before an intervention, individuals are sorted into groups based on similar growth patterns. Then the intervention effects can be compared for treated groups and control groups within the same growth group. This technique requires several years of data both before and after an intervention to establish comparison groups and then observe the intervention effects.

The regression techniques in hierarchical linear modeling or structural equation modeling are very familiar to statisticians. However, the language, equations, and coefficients in these frameworks can be difficult to understand for policy makers and educators. These techniques often involve computational packages such as HLM [54] and Mplus [55], which can act as a black box, potentially leading to misuse or misinterpretation. The techniques in this paper are designed to be more intuitive to policy makers and educators by not relying on parametric solutions with tables of computed coefficients. Nonetheless, the methods used in the literature provide a foundation and a source of comparison for the new method described in this dissertation.

2.1 Hierarchical Linear Models

In educational data, there is a natural hierarchy: students within classrooms, classrooms within schools, schools within districts, etc. The data could be disaggregated to the individual student level, however we cannot assume that the observations from students in the same classroom are independent. Alternatively, the data could be aggregated at the classroom or school level; however, this discards the within-group information, a vast majority of the data. Hierarchical models are used to capture both the student-level and group-level information. Furthermore, hierarchical models can model the repeated observations within an individual's longitudinal data.

An example of a simple two-level model, as described in Raudenbush and Bryk (2002) [47], models the mathematics performance of students with varying socio-economic status (SES) within Catholic schools or public schools. In the hierarchical form, the model has two levels; level 1 is the student level model and level 2 is the school level model. The level 2 model defines the coefficients from the level 1 model with respect to school level variables.

$$\text{Level 1 (student level)} \qquad Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + r_{ij} \tag{2.1}$$

$$\text{Level 2 (school level)} \qquad \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \tag{2.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}. \tag{2.3}$$

In the model, Y_{ij} is the mathematics score for student i in school j, X_{ij} is the SES for individual students, $\bar{X}_j$ is the average SES at school j, and W_j is a binary variable that equals one for Catholic schools and zero

for public schools [47]. By subtracting the average SES from the individual student SES, the SES is *centered* so that the intercept of the level 1 model (β_{0j}) is meaningful; instead of the intercept equaling the expected score for a student with an SES of zero, it equals the expected score for a student with an average SES. The parameters γ_{00} and γ_{01} are equal to the average score for public schools and the added average achievement for Catholic schools, respectively. The parameters γ_{10} and γ_{11} are equal to the average achievement slope with respect to SES for public schools and the added average slope for Catholic schools. The variables u_{0j} and u_{1j} are unique effects on the average scores and slopes for each school, and r_{ij} is the unique effect for individual students within those schools.

In hierarchical linear modeling (HLM) there are several assumptions pertaining to the variables in the model. First, the model must be linear in the fitted parameters (the βs and γs). Often the model is also linear in the predictor variables. This is true for most of the models using HLM, although quadratic or higher order terms can be added for the predictor variables. Second, there is an assumption of normality for the random terms and the student random terms are assumed to be independent with respect to the school variation. Specifically,

$$E(r_{ij}) = 0, \qquad \mathrm{Var}(r_{ij}) = \sigma^2, \tag{2.4}$$

$$E\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \mathrm{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00}\tau_{01} \\ \tau_{10}\tau_{11} \end{bmatrix} = \boldsymbol{T}, \tag{2.5}$$

$$\mathrm{Cov}(u_{0j}, r_{ij}) = \mathrm{Cov}(u_{1j}, r_{ij}) = 0. \tag{2.6}$$

HLMs are often *mixed* models, meaning they contain a combination of *fixed* and *random* effects [46]. Fixed effects are effects that are common for everyone or within groups and they are typically used for categorical variables (race, SES, etc.). The mean values, such as the γs from the above model, are fixed effects. Random effects are fluctuations that are unique for each individual or group and they are typically used for variables with many values (students, schools, etc.); u_{0j}, u_{1j}, and r_{ij} are examples of random effects. Often researchers are more concerned with the fixed effects (aka the model for the means). The random effects (aka the model for the variance) are usually oversimplified with assumptions of normality, independence, homoscedasticity, etc. This may not always accurately represent the variance in real datasets.

For longitudinal data, a two-level HLM would represent the within-person and between-person components of the model as the two levels; within-person changes are expressed in level 1, and level 2 contains the between-person differences. An example of a simple linear two-level longitudinal HLM is [46]

$$\text{Level 1 (individual growth)} \qquad Y_{ti} = \beta_{0i} + \beta_{1i} a_{ti} + e_{ti} \tag{2.7}$$

$$\text{Level 2 (between person)} \qquad \beta_{0i} = \gamma_{00} + u_{0i} \tag{2.8}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}, \tag{2.9}$$

where Y_{ti} is the observed outcome for person i at time t, a_{ti} is the age (or other time-related variable), γ_{00} and γ_{10} are fixed effects of the intercept and slope, u_{0i} and u_{1i} are random unique effects of the intercept and slope, and e_{ti} is the unique random term for individuals at each observation. There could

Fixed Effect	Coefficient	se	t Ratio	
Mean initial status, γ_{00}	-0.135	0.005	-27.00	
Mean growth rate, γ_{10}	0.182	0.025	7.27	
Random Effect	Variance Component	df	χ^2	p Value
Initial status, u_{0i}	1.689	139	356.90	<0.001
Growth rate, u_{1i}	0.041	139	724.91	<0.001
Level 1 error, e_{ti}	0.419			
Reliability of OLS Regression Coefficient Estimate				
Initial status, β_{0i}	0.854			
Growth rate, β_{1i}	0.799			

Table 2.1: Results from a two-level longitudinal HLM measuring natural science knowledge. Replicated from Raudenbush and Bryk (2002), p.165 (adjusted notation). *se* is standard error, *df* is degrees of freedom, *OLS* is ordinary least squares.

also be level 2 parameters that could measure the contribution of individual characteristics (e.g. ethnicity, SES) or experimental treatment (e.g. course-taking, tutoring) [47], in which case the level 2 equations would look like those in the Catholic/public school example above and W_j would represent that characteristic or treatment variable.

The results from an HLM are presented in the literature in tables of coefficients, p values, etc. An example of the results from a linear mixed model, such as the longitudinal example above, is shown in Table 2.1. While statisticians are very adept at interpreting these results tables, they may not be easily interpreted by educators or policy makers without expertise in HLM. Thus, the method used in this dissertation employs visualizations to more easily interpret the qualitative and quantitative results. These visualizations

are not a substitute for statistical tests although they are a valuable starting point.

The TEA uses a simple two-level HLM to measure students' TAKS score progress through the Texas Projection Measure (TPM). The composite equation (substituting the level 2 equations into the level 1 equation) is [2]

$$TAKS_{ij} = \gamma_{00} + \gamma_{10}(TAKS\_M_{ij}) + \gamma_{20}(TAKS\_R_{ij}) \\ + \gamma_{30}(School\_Mean_j) + u_{0j} + r_{ij}, \tag{2.10}$$

where $TAKS_{ij}$ is the TAKS score for student i in school j for the subject of interest, $TAKS\_M_{ij}$ is the mathematics score, $TAKS\_R_{ij}$ is the reading score, $School\_Mean_j$ is the mean score for the subject of interest at the school, the γs are fixed effects, u_{0j} is a random intercept school effect, and r_{ij} is a random individual effect. By using mostly fixed effects, the model does not take into account any latent factors that could affect performance at the school or student level outside of a random component. Further documentation about the TPM has been removed from the TEA website.

2.1.1 Individual Growth Models

When studying the change of outcomes over time, the within-person component of the hierarchical model is of particular interest. This part of the model is sometimes called the individual growth model. This technique is particularly useful for smaller sample sizes. Empirical growth records show the exact observed outcomes of individuals over time. These are fitted with linear,

quadratic, or other parametric expressions that are called *trajectories* [48]. In particular, these trajectories or individual growth curve models are used to estimate the between-person differences in within-person growth [56].

Individual growth models are often mixed models, with fixed and random effects. For linear trajectories, the intercepts and slopes may contain fixed and/or random effects, so that the parameters describing the linear functions vary between individuals. Individual growth models provide more flexibility over traditional methods like repeated analysis of variance (ANOVA) because the observations do not need to be at the same time for each individual and the growth need not be linear.

While HLM results are typically presented using tables of coefficients, the individual growth models are often presented using plots, either for single individuals or a collection of people. As a result, this technique is not often implemented with large numbers of participants. The average change trajectory can be plotted, representing the best fit for the average outcome variable over time. Furthermore, the individual and average change trajectories for groups can be compared to show the differences in growth for the groups.

2.1.2 Structural Equation Models

There is a very similar but distinct family of statistical methods called structural equation models (SEM). The distinction is somewhat superficial as the results from the methods are nearly identical, although they do emphasize different components of the model. The structured equations in the model

are almost always in the same form as HLMs. There are, however, two major attributes of SEM that highlight its unique contribution: path diagrams, and latent curve modeling.

The relationships between the variables in the model are expressed through path diagrams. These diagrams use shapes to convey the types of variables (e.g. observed, unobserved, constant) and directed arrows to convey the direction of assumed causality, pointing from the independent variable to the dependent variable. While HLM focuses on the quantitative impact of the relationships, SEM emphasizes causality. The path diagram corresponding to the following equation is shown in Figure 2.1:

$$Y_n = \beta_0 \cdot 1 + \beta_n \cdot X_n + e_n \cdot 1. \tag{2.11}$$

The interpretation of path diagrams is not intuitive to people unfamiliar with SEM. Computer programs such as Mplus and LISREL are used to compute the parameters and draw the path diagrams [49].

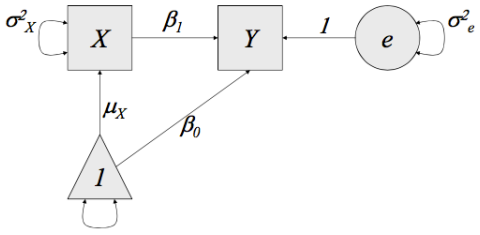

Figure 2.1: Replicated path diagram [6] corresponding to Equation 2.11.

The other major contribution by SEM, which is sometimes included in HLM, is the use of *latent variables*. Latent variables are unobserved but the

effects of the latent variables can be seen in the observed variables or indicators. One example of a latent variable is the true score from Classical Test Theory. Intelligence or academic performance are latent variables that could be measured through test scores, IQ, etc. Random terms are also technically latent variables. The relationships between latent variables and observed variables in SEM are made explicit through a measurement model. A structural model imputes the relationships between the latent variables [57]. Since the latent variables are often the outcomes of interest, SEM frames the problem in a way that emphasizes these underlying traits.

2.1.3 Grouped Multilevel Models

In both the HLM and SEM frameworks, individual growth modeling or latent growth modeling can be extended to groups of individuals. In HLM this is called group-based trajectory modeling and in SEM it is called latent class growth modeling. The individuals are not arbitrarily placed into groups, rather the groups are formed from the data to combine individuals with similar growth patterns into the same group.

In these grouped growth models, the total population is assumed to be a mixed population of individuals that could be characterized into distinct groups or classes by a time-invariant latent variable [58]. The probabilities of belonging to each group are calculated for each individual and then they are placed into the most likely group. The average growth trajectory for each group is then calculated. It is important to note that the longitudinal growth

for each individual is considered before placing them into groups, not just initial conditions. The computation of group membership therefore requires computer software such as SAS [53]. Daniel Nagin, the primary developer of group-based trajectory modeling and latent class growth modeling, has used this technique to study juvenile delinquency [52, 58] and disease biomarkers [59].

These group growth modeling techniques are also powerful tools for studying the effects of treatments or interventions. In particular, the trajectory groups could be established using longitudinal data prior to the treatment. Each group could be divided randomly, with half receiving the treatment and the other half acting as a control. The differences in development between the treated and control groups within the same initial trajectory can be observed over a period of time after the treatment. By using the grouped trajectories to establish similar prior development, the treatment and control groups are more appropriately established for longitudinal comparisons.

2.2 Age-Period-Cohort Effects

Between-person and within-person changes might be observed as any of three time-related variations: age effects, period effects, and cohort effects. Age effects represent changes related to aging although in the context of education, age effects could instead be thought of as grade effects, changes that occur during the progression of a student through school. Period effects represent changes occurring during a specific time-period, affecting people of all

ages similarly. Cohort effects represent formative experiences, changes that are unique to people who experience the same events at the same time, often because they were born in the same time-period. In an education context, cohort effects may relate to changes that affect a graduating class of students, students who progress through the grades together.

Longitudinal student data could be organized in a table by grade and school year, as in Table 2.2, to help clarify grade, period, and cohort effects. Each grade level is represented by a row of the data; therefore, to control for grade effects (by ignoring them), one would use the data from a single row. Each column represents a school year; therefore, a single column of data controls for period effects. A diagonal line, going down and to the right, represents a graduating cohort of students. The cohort of students that graduated in 2012 is highlighted in Table 2.2 as an example. The students in a cohort progress one grade per school year, which is the traditional student pathway. If a student skips a grade or is retained in a grade then the student moves to a different cohort. Selecting the data along a diagonal would control for cohort effects by only observing the data for a single cohort.

It can be difficult to separate age, period, and cohort effects from each other. This is due to the inherent relationship between age, time, and cohort. For studies using birth cohorts—people born in the same year—the relationship between age, year, and birth cohort is $Year - Age = Birth\ Cohort$. For students in primary or secondary school, the relationship between the year (spring semester), grade, and cohort (labeled by the year of high school grad-

	2002-2003	2003-2004	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009
Grade 3							
Grade 4							
Grade 5							
Grade 6							
Grade 7							
Grade 8							
Grade 9							

Table 2.2: Grade-Period table: each grade is represented by a row, each period (school year) is represented by a column, and each cohort is represented by a diagonal (highlighting the cohort of students that graduated in 2012).

uation) is $Year - Grade + 12 = Cohort$, assuming the students follow the traditional path of one grade per year. These relationships make it difficult to design a study to isolate the age/grade, period, or cohort effects because it is difficult to control for more than one of these effects. In addition, there may be several concurrent influences, causing a combination of age/grade, period, and cohort effects. Educational policy changes are usually either period effects or cohort effects, depending on the process of implementation and the intended recipients.

Selecting data in a single row, column, or diagonal in a dataset organized like Table 2.2 sets the grade, school year, or cohort at a constant value, thereby controlling for its effects. However, this selection simultaneously confounds the effects of the other two types. For example, if we were to select data from a *cross-section* in time, represented by a single column, then changes in the results along a column could either be due to the differences in grade

level or the differences in cohorts. Similarly, if we were to select data along a diagonal, following a single cohort longitudinally, it would be impossible to distinguish if any developmental changes happened as a result of the grade level or the school year.

To tease apart the confounded effects in a single row, column, or diagonal, it helps to compare the results to a different row, column, or diagonal. When comparing the results from different years (columns), for example, if the relationship between the outcome and the grade level is the same for every year, then that effect is likely a grade effect. If the relationship changes, then it may be harder to identify the effect as a grade, period, or cohort effect. Prior information may help to identify the type of effect. For example, the Student Success Initiative was a cohort effect, since the implementation occurred gradually with the cohort of 2012 (and also affected later cohorts).

Age-period-cohort (APC) analysis is a popular tool in the social sciences and in medicine. Epidemiologists assemble similar tables to Table 2.2 with morbidity or mortality rates for a disease with respect to the period and age of the patients [60]. Sociologists Robinson and Jackson used APC analysis to study the declining interpersonal trust between Americans before September 11th [61]; Clark and Eisenstein showed with APC analysis that the decline in trust continued through 2013 [62]. In most APC analysis, the researchers use a linear regression equation:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{a-i+j} + \epsilon_{ij} \tag{2.12}$$

where Y_{ij} is the observed rate of the outcome variable, μ is the mean, α_i is the fixed effect of the ith age category, β_j is the fixed effect of the jth period category, γ_{a-i+j} is the fixed cohort effect for period j and age i (out of a age categories), and ϵ_{ij} is the random term with a null expectation value [60]. *Random* age, period, or cohort effects are also used in some of the literature [63].

2.2.1 Cross-Sectional Models

To minimize or isolate the potential sources of time-related change, studies may focus on a single time-period or a single cohort. Studies that use data from one time-period are known as cross-sectional models. By using only one time-period of data, period effects are removed (ignored) from the analysis and the time required for data collection is minimal, an attractive feature for costly studies. Cross-sectional models show the range of outcomes within a moment in time and can be helpful for identifying between-person variation. The students within a cross-section can be aggregated into a *synthetic cohort*, which represents students at every grade throughout school. If there were no cohort (or period) effects, a synthetic cohort would accurately identify age/grade effects. However, without other time-periods for comparison, age/grade effects and cohort effects are completely confounded. Extending the study longitudinally can help to separate the age/grade effects from the cohort effects, although this could reintroduce period effects.

One traditional method of extending cross-sectional techniques to longitudinal data, especially in medicine [64] but also in education research [65],

is to use repeated measures of ANOVA to study the change in the average outcome for a group over time. In one-way ANOVA, for example, the variability between groups is compared to the variability within groups (for repeated ANOVA a group might be comprised of measurements within a cross-section) by using the F-statistic:

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}\ , \tag{2.13}$$

where n_i is the number of observations in group i (out of k groups), $\bar{Y}_i$ is the average in the group, $\bar{Y}$ is the total average, Y_{ij} is the jth observation in group i, and n is the overall sample size. For ANOVA, the independent variables must be categorical, and in the case of repeated ANOVA, one of the independent variables represents the time periods of measurement. ANOVA assumes random samples with normally distributed observations and common variances. While repeated ANOVA is ideal for quickly determining the differences in average outcomes for groups over time, it cannot show variation in individuals over time. It is also ill equipped to study non-linear effects. As a result, many researchers limit the use of repeated ANOVA to a first attempt at analyzing the effects, and use general linear mixed models to more rigorously analyze the data [66].

2.2.2 Single-Cohort Models

Single-cohort studies use the longitudinal data from a single cohort to study within-person change [67]. Cohort studies are helpful for studying the evolution of individuals within the cohort, especially considering that a person may have many experiences that build upon each other to inspire a particular growth pattern. Single-cohort studies require several data points for each individual, which can take years of data collection and can lead to attrition. And, while cohort studies remove cohort effects from the analysis (by ignoring them), age effects and period effects are confounded. By comparing outcomes from multiple cohorts, age effects and period effects can be separated, while possibly reintroducing cohort effects. Nonetheless, cohort studies are a popular method to study developmental changes in longitudinal data.

Conaway, Keesler, and Schwartz, state education practitioners in Massachusetts, Michigan, and Tennessee respectively, published a paper about the potential of rigorous longitudinal studies, such as cohort studies, for influencing policy decisions and informing educators [68]. They described the unique ability of State Longitudinal Data Systems (SLDSs), like the dataset at the ERC, to provide complete pictures of the students' educational paths and establish causality in intervention studies. Longitudinal studies are particularly useful for determining the long-term consequences of interventions. It is long-term outcomes that interest policy makers, since high school graduation rates and college readiness eventually should lead to greater economic competitiveness [69, 70, 71].

However longitudinal studies face some fundamental difficulties. One of these difficulties is due to the mismatch between the frequency of political decisions and the duration of a student's education, and thus the length of a longitudinal study. Particularly for interventions affecting younger students, for a longitudinal study to measure the impact of the intervention on high school graduation rates, it would take about a decade for the affected students to progress through school. On the other hand, educational policy tends to change about every two years [70] . Regarding longitudinal studies, Conway et al. say, "these studies by themselves are not particularly responsive to the way business gets done...we want to know not only whether the initiative is working but also how it might work better...within the relatively short period needed to establish political buy-in" [68].

Therefore, longitudinal cohort studies may not be the ideal tool for studying the influence of educational policy changes. Accelerated longitudinal models may provide a better alternative, which decreases the duration of the study while providing longitudinal results to study within-person changes.

2.2.3 Accelerated Longitudinal Models

Accelerated longitudinal design (ALD) studies, also known as cross-sequential design studies, are a compromise between cross-sectional studies and longitudinal studies [72]. ALDs use data from multiple overlapping cohorts beginning at different ages to span a large age range while using only a few years of data. For example, Miyazaki and Raudenbush used the National

	Grade									
	3	4	5	6	7	8	9	10	11	12
Cohort 2013	2003-2004	2004-2005	2005-2006	2006-2007						
Cohort 2011			2003-2004	2004-2005	2005-2006	2006-2007				
Cohort 2009					2003-2004	2004-2005	2005-2006	2006-2007		
Cohort 2007							2003-2004	2004-2005	2005-2006	2006-2007

Table 2.3: Example of data used in an ALD with four cohorts, covering ten grades in only four years.

Youth Survey, which contained data for 7 adjacent cohorts over 5 years, to study the development of antisocial attitudes from ages 11 to 21 [73]. ALDs study growth over a large age/grade range without needing to wait the full time period as in longitudinal studies. This reduces the cost of the study as well as the attrition due to missing data while producing results in a time period that allows for more political influence.

An example of the data that might be used in an ALD is shown in Table 2.3. In this example, data is used from four cohorts, following each for four years. By using data from overlapping cohorts in this way, the full duration of the study lasts only four years while the data represents information from ten grade levels. If this study had used a single cohort, this would have taken six extra years to complete the study.

ALDs are modeled using HLM, usually linear mixed models with fixed or random effects for each cohort [74]. In the hierarchy, longitudinal observa-

tions are nested within individuals who are nested within cohorts. ALDs are particularly useful for identifying cohort effects. Studies have been done to investigate the ideal number of cohorts, years, and overlapping years depending on the size of the dataset [75].

John Mirowsky and Jinyoung Kim [76, 77] combine ALD with vector graphs, exploring changes in depression over time in a novel way. They define aging vectors, which use an HLM to express the relationship between the outcome, age, and follow-up time. One vector might have the form:

$$Y_{it} = a_i + b_i t + e_{it} \tag{2.14}$$

$$a_i = a_0 + a_1(A_{i0} - k) + a_2(A_{i0} - k)^2 + u_{ai} \tag{2.15}$$

$$b_i = b_0 + b_1(A_{i0} - k) + u_{bi}, \tag{2.16}$$

where Y_{it} is the outcome variable for person i at time t and A_{i0} is the age at the midpoint of the follow-up which is centered on a reference age k. Therefore $(A_{i0} - k)$ is the cohort index number with respect to a reference cohort. The model has mostly fixed effects except for the random terms e_{it}, u_{ai}, and u_{bi}. By having t in the within-person level of the model instead of A_{i0}, the level 1 model describes the changes in a follow-up period rather than changes with aging, thus allowing for more flexibility when using multiple overlapping cohorts with different age ranges [76]. This HLM allows for predictions of changes in the outcome with respect to the cohort and the follow-up time. The aging vectors are computed for each cohort in the ALD, creating a series of vectors over the full age range.

Mirowsky et al. compare the aging vectors to cross-sectional curves and synthetic cohort trajectories. Cross-sectional curves simply plot the outcome with respect to age during one period. Synthetic cohort trajectories link aging vectors head-to-tail by simply shifting the vertical position of the aging vectors so that the vectors are connected. Despite changing the initial outcome value for each segment, they do not adjust the slope of the arrows. In the absence of cohort effects, the cross-sectional curve and the synthetic cohort trajectory are the same.

When there are cohort effects, Mirowsky et al. express this through a trend function T_i, which gives the difference between period effects and age effects as a linear function of the cohort index:

$$T_i = \hat{b}_i - \frac{\partial \hat{a}_i}{\partial A_{i0}} \tag{2.17}$$

$$T_i = (b_0 - a_1) + (b_1 - 2a_2)(A_{i0} - k) \tag{2.18}$$

The trend function is essentially the difference between neighboring cohorts. Mirowsky defines a virtual cohort projection which uses the trend function to find the implied trajectories of cohorts with respect to a reference cohort. There is an assumption that the age-specific trend is the same between every neighboring cohort. The virtual cohort projection for a cohort d years older than the reference cohort is then defined as:

$$V(k+d) = \hat{a}_{i0} + T_i(A_{i0} - (k+d)) \tag{2.19}$$

$$= V(k) - dT_i \tag{2.20}$$

This gives an estimated trajectory for each cohort in the ALD. The virtual cohort projection might be an improvement over the observed longitudinal trajectories since attrition can lead to biased results.

Cross-sectional curves, synthetic cohort trajectories, and virtual cohort projections each represent a different aspect of the age-period-cohort effects, though Mirowsky does not consider period effects in his study of depression (he considers any period effects as survey-year residuals that are uncorrelated with age or cohort) [77]. A cross-sectional curve shows the longitudinal trajectory if there are no cohort (or period) effects (the function of outcome with respect to age is constant over time). A synthetic cohort trajectory shows the longitudinal trajectory if there is no interaction between age and cohort (the cohort effects are independent of age). A virtual cohort projection shows the longitudinal trajectory if the age-specific differences between cohorts remain the same over time.

While there are many methods in the social sciences and medicine that analyze within-person and between-person changes over time as well as any age, period, or cohort effects, the majority of the methods use parametric (often linear or quadratic) expressions for the outcome variable over time. For many popular research areas, the relationships between variables are established in the literature without many alternatives. Residuals are almost always assumed to be normally distributed. Additionally, the results are usually presented as tables of coefficients, which are not particularly accessible to readers without a statistical background. The papers published by Marder

and Bansal (2009) and Bendinelli and Marder (2012) take a different approach, which requires fewer assumptions and presents the results through visualizations. These papers provide a foundation for the technique developed in this dissertation.

2.3 Foundation Techniques

In their 2009 paper [7], Marder and Bansal applied fluid flow modeling techniques from statistical mechanics to analyze student standardized test scores. In the abstract, the students' scores were conceptualized to be flowing through the grades or years similarly to a fluid, with random individual score fluctuations mirroring the random behavior of individual particles. In particular, Marder and Bansal used the Fokker-Planck equation, which is a partial differential equation that describes the evolution of the probability density function for the velocity of a particle experiencing drag or random forces.

In one dimension, the Fokker-Planck equation takes the form

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)], \tag{2.21}$$

where p is the probability density, μ is the drift coefficient and D is the diffusion coefficient. The drift captures the forward flow of the fluid and the diffusion captures the random movement. In the case where the drift is zero and the diffusion is constant, for example, the Fokker-Planck equation describes Brownian motion. Marder and Bansal developed an expression for the change in the number of students in a score bin that is inspired by the Fokker-

Planck equation (the notation has been changed to match the notation used by Bendinelli [78]).

If $N_{t,g,k}$ is the number of students in year t and grade g whose score fell within bin k, an expression for the change in the number of students in the bin as the students progress to the next grade is [78]

$$N_{t+1,g+1,k} - N_{t,g,k} = -\frac{\partial}{\partial k}[v_{t,g,k}N_{t,g,k}] + \frac{\partial^2}{\partial k^2}[D_{t,g,k}N_{t,g,k}] - \Delta_{t,g,k}, \qquad (2.22)$$

where v is the *velocity* of the students' scores and D is the diffusion coefficient, given by

$$v_{t,g,k} = \frac{\sum_{\alpha=1}^{N_{t,g,k}} (s_{t+1}^{\alpha} - s_t^{\alpha})}{N_{t,g,k}} \qquad (2.23)$$

$$D_{t,g,k} = \frac{1}{2}\frac{\sum_{\alpha=1}^{N_{t,g,k}} (s_{t+1}^{\alpha} - s_t^{\alpha})^2}{N_{t,g,k}}, \qquad (2.24)$$

and Δ is a loss term that compensates for missing scores and is equal to the number of students who had a score in year t but not in $t+1$ subtracted by the number of students who had no score in t but had a score in $t+1$. In the velocity and diffusion terms, s_t^{α} is the score of student α in the year t.

In addition to developing this theory of student scores, Marder and Bansal used flow plots to visualize the flow of student scores through the space of score and grade. As seen in Figure 2.2, the x-axis shows the grade transitions and the y-axis shows the raw scores in percentages with respect to the maximum scores. In each grade, the students are grouped into score bins

determined by the score percentages (ten bins total) and for each group, the average change in score between that year and the next, when the students have progressed to the next grade, is plotted using arrows. The slope and size of the arrows represent the average change in score and the number of students in the bin, respectively. The background is shaded to display the score cutoffs established by the TEA.

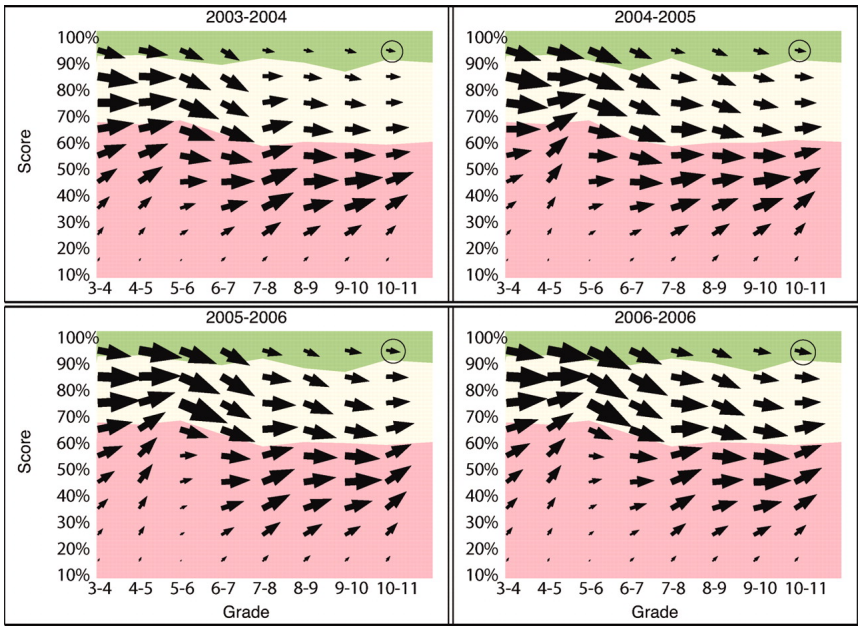

Figure 2.2: Flow plots for low-income students between 2003 and 2007 representing score changes for each grade and score bin. The tan band surpassed the Met Standard cut-off score, the green band achieved Commended Performance [7].

These flow plots were the foundation of the techniques developed by Bendinelli and Marder in their 2012 paper. The concept of using statistical and fluid mechanics techniques to analyze student scores was expanded with the use of trajectory and streamline plots. The alternatively binned streamlines

developed in this dissertation directly build off the foundation established by Bendinelli and Marder.

2.3.1 Trajectory Plots

While the social sciences have developed several methods that are called trajectories (individual growth models, group-based trajectory modeling, etc.), the trajectory plots developed by Bendinelli and Marder are inspired by physics and the movement of objects over time. In the context of student test scores, the social models and the physical models are very similar, representing the flow of scores over time. The main difference between these models is that the social models are almost always parametric and often linear, whereas the physical trajectories are not necessarily parametric and allow for nonlinearity. In the rest of the dissertation, *trajectories* and *trajectory plots* will be used to represent the technique developed in Bendinelli and Marder (2012).

Trajectories represent the change of an outcome variable over time. Trajectory plots, such as those shown in Figures 2.3, 2.4, and 2.5, show the average scores for a group of students over time. I group the students into bins determined by their score in the initial exam (3rd grade). The students remain in these groups throughout the entirety of the longitudinal analysis for as long as they have documented scores (and assuming they have at least the first two years of data). Therefore, trajectories represent the change in average scores over time, with respect to the initial score. There are approximately 250,000 students included in the analysis of a single cohort.

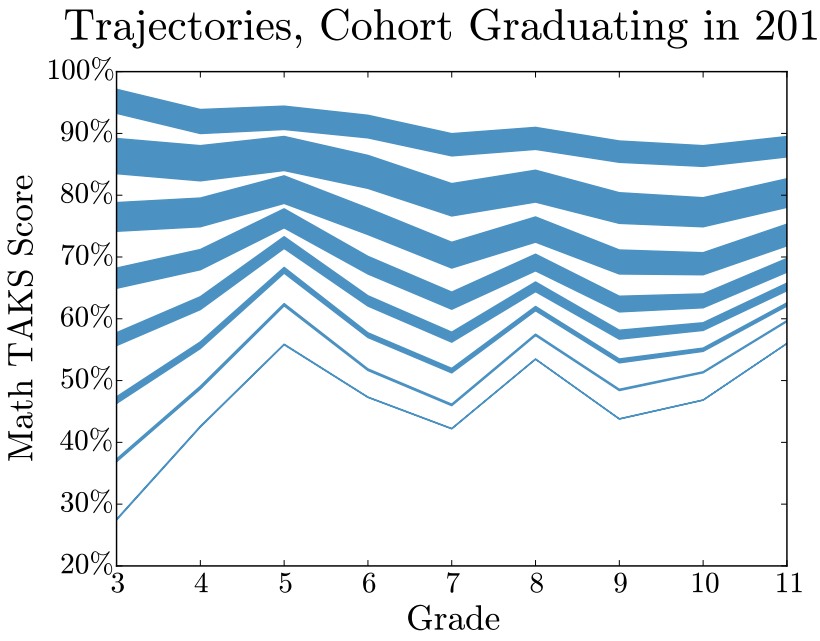

Figure 2.3: Trajectories for the cohort of students who graduated in 2012, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group.

To be more specific, all of the 3rd graders in a certain year are placed into bins determined by their percent score (90-100%, 80-90%, 70-80%, etc.) on the mathematics TAKS exam. Students can also be further disaggregated by demographic variables, course-taking, or other variables (explored in Chapter 4). Once the groups are formed, the average scores for those groups are calculated in 3rd grade and the subsequent years and grades. These average scores are then connected with linear segments to create the trajectory (although the segments need not be linear). The students remain in the trajectories for as long as they stay in the cohort, following the traditional path

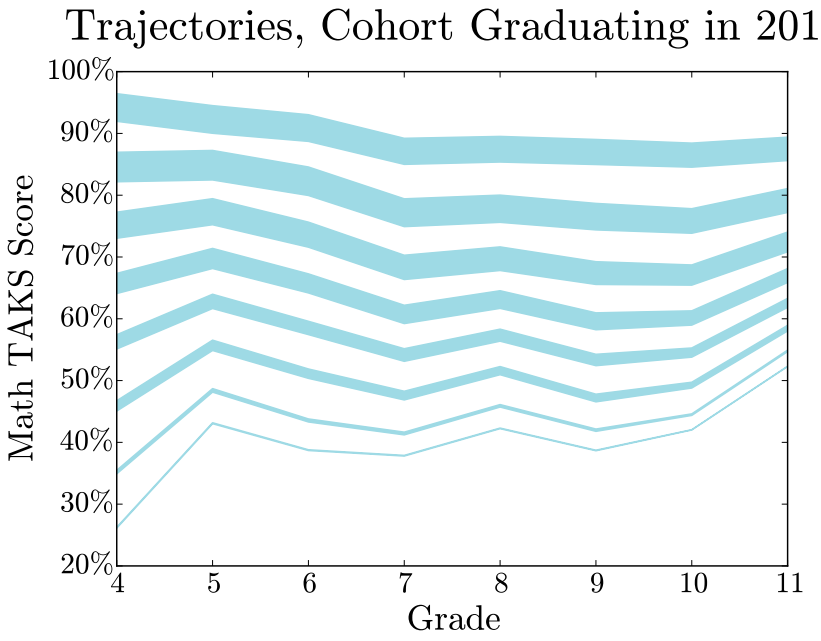

Figure 2.4: Trajectories for the cohort of students who graduated in 2011, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group.

of progressing one grade each year. Students are not included in the trajectory if they join the cohort after 3rd grade because they cannot be sorted into a group. Trajectories are unable to capture the performance of non-traditional students who join or leave the cohort.

This form of analysis is similar to group-based trajectory modeling. In group-based trajectory modeling, student groups are determined by similar growth patterns established the several years prior to the treatment. For each group, parametric functions describing the time-variance of the outcome variable are fitted to the student scores [52, 58]. Bendinelli's trajectory plots

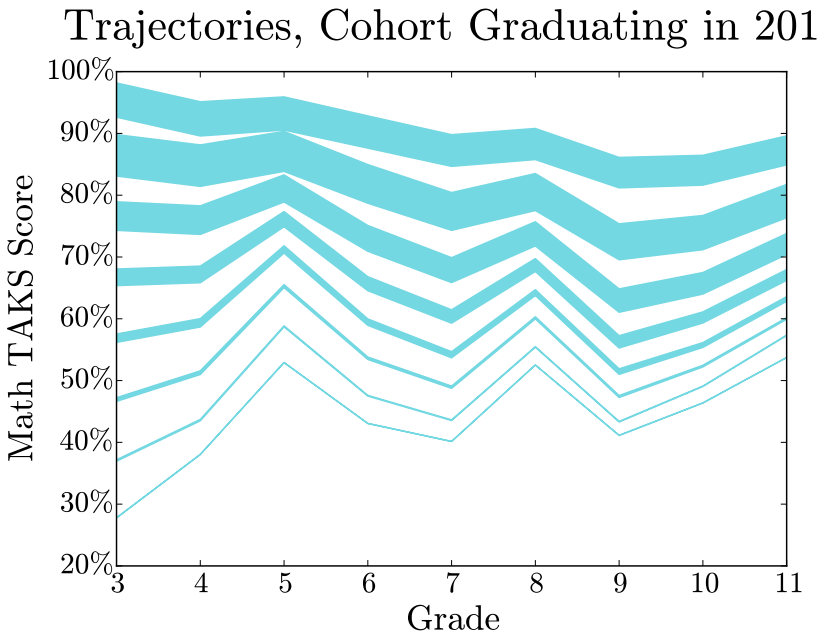

Figure 2.5: Trajectories for the cohort of students who graduated in 2013, representing the average score over time for students grouped by their 3rd grade score. The thickness of the trajectory is proportional to the number of students in that group.

only use the scores from the initial exam to determine the groups. This is a simpler way of establishing groups that is intuitive to educators and policy makers. It also lends itself to answering simple questions, such as "what scores can a student expect to get if they start out with score a?" The major downfall of using only one exam to group students is that the scores are noisy measures of the students' abilities, and the random fluctuations result in regression to the mean, which will be discussed in depth in the following chapter.

Trajectory plots are of fundamental interest because they track the average performance of a cohort of students at all grades and performance levels

exactly. Therefore, trajectories provide an accurate depiction of longitudinal student performance on a large scale and they can be used to analyze the outcomes of policies in the long term. However, policies typically have a shorter duration than the nine years of data that it takes to construct a full trajectory, and so the policy has likely already changed by the time the results can be analyzed with this method. In addition, attrition can introduce bias to the sample and becomes more of an issue with longer studies. It is therefore necessary to identify other techniques that permit more timely analysis.

2.3.2 Streamline Plots

Streamlines provide an approximate way to find trajectories and are a widely used technique in fluid mechanics. To calculate test score streamlines, a vector field is used to represent the average change in score for each score bin and grade. This vector field is visualized by the flow plots developed by Marder and Bansal. Streamlines are constructed as integrals of the vector field, representing the flow of student scores. If student score changes were completely deterministic functions of their scores, and if the educational environment did not change over time, then streamlines and trajectories would be identical.

In fluid mechanics, a streamline is an instantaneously tangential curve to a velocity vector field that represents the movement of a particle in a fluid [79]. Using this analogy, we can define a *velocity* for a group of stu-

dents as the average change in score between two consecutive years:

$$v = \frac{1}{N} \sum_{\alpha} (s_{t+1}^{\alpha} - s_t^{\alpha}),$$

where v is the average change in score for a group of students, N is the number of students in that group, and s_t^{α} is the score of student α in the year t. This is the same definition used in the Fokker-Planck-inspired equation developed by Marder and Bansal. Similar to the trajectory plots described above, the students can be sorted into bins according to their percent scores in a particular grade. Once the groups are formed, velocities can be calculated for each group, equaling the average change in score from that grade to the next grade (in the following year). This would form a column of vectors, with slopes equaling the average change in score between the two grades, one vector for each score bin. This process can be repeated in each subsequent grade, regrouping the students each time and then calculating the corresponding velocities. Together these make up the vector field, which represents the velocity for students in each grade and score bin.

For each grade transition, a continuous function relating the score in one grade to the anticipated change in score to the next grade can be interpolated from the velocity calculations, filling in the space between the arrows. A streamline is constructed by starting with an initial score in 3rd grade and then using the interpolation function for 3rd grade to find the anticipated score in 4th grade. This new score is then used in the function for 4th grade to find the anticipated score in 5th grade. This process is repeated for each

grade, using the anticipated score from the previous calculation as the initial score in the next velocity calculation. As a result, a piece-wise linear streamline strings together the changes in score from each grade to represent a sequence of anticipated scores. In essence, streamlines are constructed from interpolated changes in score.

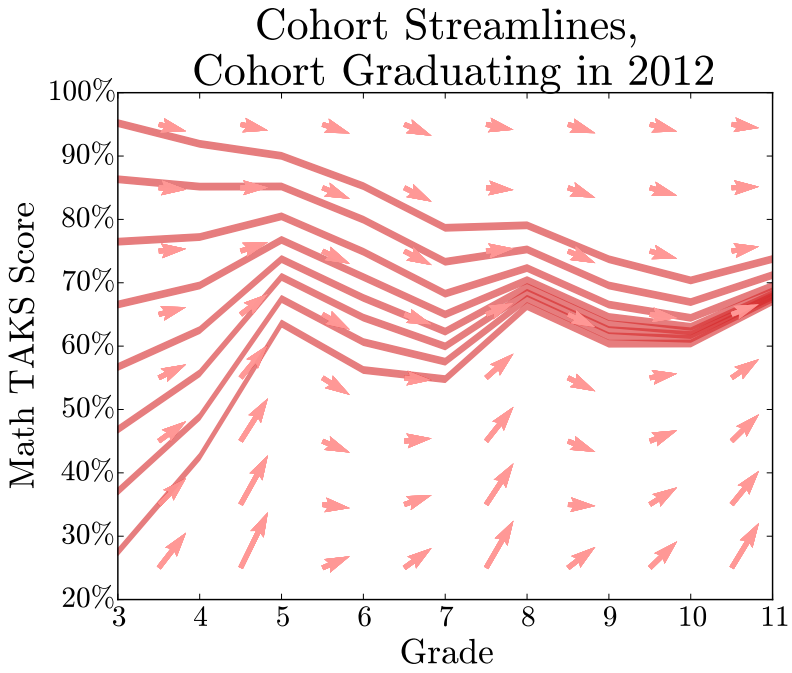

Figure 2.6: Arrow plot and corresponding streamlines for the cohort of 2012. The arrows show the change in score by grade and score bin. The streamlines show interpolated scores over time based on the arrows.

There are two options for timing conventions in a streamline plot. One option is a cohort streamline (Figure 2.6), which follows the students longitudinally as they progress through school, each exam taking place in a different year. In a cohort streamline plot, the overall set of students remains mostly

the same, and these students are re-sorted into new score bins each grade. The other option is a snapshot streamline, which is derived from the changes in score for a synthetic cohort of students during two consecutive cross-sections of time. This synthetic cohort is comprised of students from each grade in one year. The students from each grade are grouped into score bins for that year, and the velocities are calculated for those groups from that year to the next. A snapshot streamline represents the flow of scores through the grades during two consecutive years. Figure 2.7 shows the snapshot streamline for 2003-2004, sorting the students by their 2003 scores and calculating the changes in score between 2003-2004.

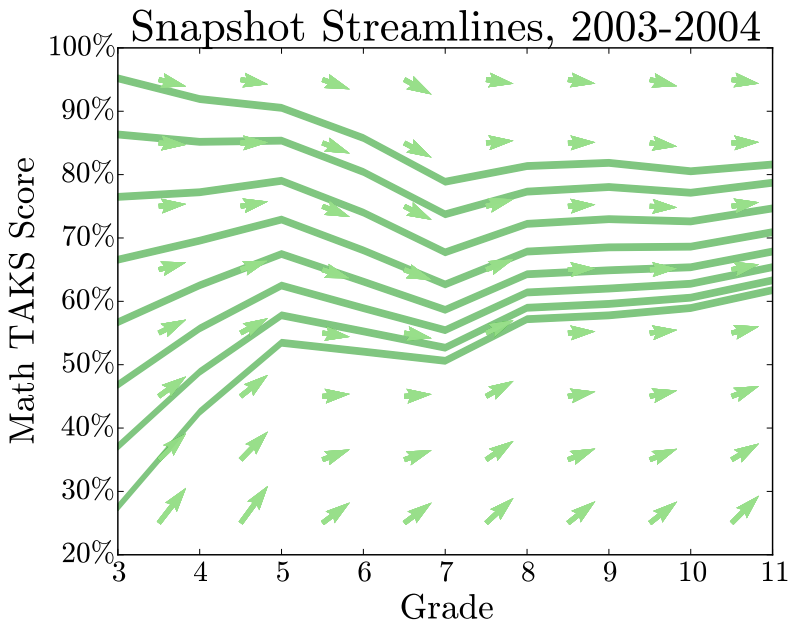

Figure 2.7: Arrow plot and corresponding streamlines for each grade transition between 2003 and 2004. The arrows show the change in score by grade and score bin. The streamlines show interpolated scores through the grades based on the arrows.

	Sorting Grades	Average Score Change	Study Duration	Study Design
Trajectories	Grade 3	Observed	Nine Years	Single Cohort
Streamlines	Every Grade	Interpolated	Two Years	ALD

Table 2.4: Summary of the differences between trajectories and snapshot streamlines.

A summary of the main differences between snapshot streamlines and trajectories can be seen in Table 2.4. Students are sorted into score bins once in trajectories, but every grade in streamlines. The average change in score which is represented in each segment uses the observed scores in trajectories and the interpolated scores in streamlines. The duration of the trajectory data is nine years, the duration of the streamline data is only two years. Trajectories require that students have a 3rd grade score (to be sorted), and the students will remain in the analysis for as long as they stay in the cohort (they follow the traditional path). Streamlines require that the student have scores in any two consecutive grades during the duration of the study and they have an ALD.

Both cohort streamlines and snapshot streamlines have unique uses. Cohort streamlines are most comparable to the trajectory plots because they both follow a cohort of students longitudinally. Therefore, the accuracy of the streamline method can be determined by comparing the results of a cohort streamline to a trajectory plot for the same cohort. Snapshot streamlines cut down considerably on the years of data required to perform the complete analysis from 3rd to 11th grade, shortening the duration from nine years to two.

Thus snapshot streamlines potentially provide a quick method for approximating longitudinal data. In addition, snapshot streamlines from different periods can be compared to identify the effects of intermediate interventions.

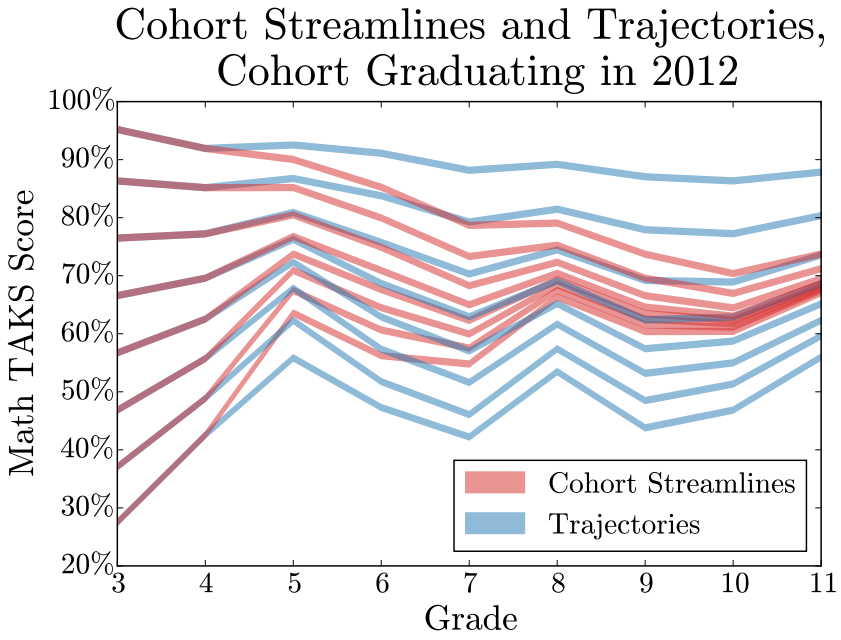

Figure 2.8: Cohort streamlines and trajectories for the cohort of 2012. The convergence of the streamlines is due to regression to the mean.

In practice, when the cohort and snapshot streamlines are computed and compared with trajectories, one finds that they deviate considerably. As seen in Figures 2.8 and 2.9 the streamlines converge towards one another, especially when compared to the trajectories. This convergence results from regression to the mean. Without addressing regression to the mean, the streamline plots cannot be used to accurately represent longitudinal student scores.

In the next chapter, a simple theory will be used to describe the ex-

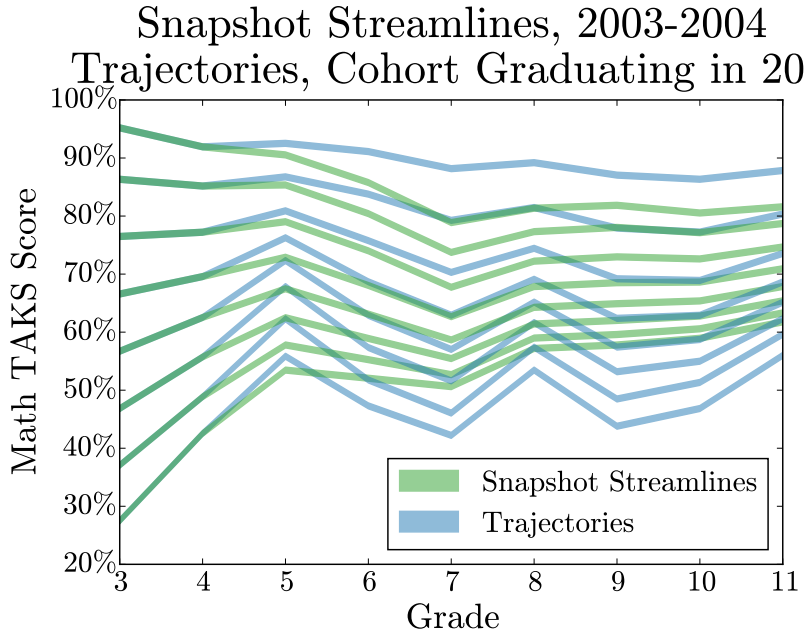

Figure 2.9: Snapshot streamlines for 2003-2004 and trajectories for the cohort of 2012. The convergence of the streamlines is due to regression to the mean.

tent of regression to the mean in trajectories and streamlines. A new method of binning the students that reduces regression to the mean, called alternative binning, will be described. The three methods—trajectories, streamlines, and alternatively binned streamlines—will be compared to demonstrate the benefits of using alternatively binned streamlines.

Chapter 3

Methods

3.1 Regression to the Mean

Regression to the mean (RTM) is a statistical consequence of the random fluctuations within observed measurements and it affects research in many areas, including sports [80], economics [81], and education [82]. Within repeated measures of a variable containing a random component, RTM describes the tendency for extreme values within a population to become less extreme. RTM is dependent on the joint distributions of the variable during the repeated measures and on the magnitude of the random fluctuations. The effects of RTM are particularly pronounced when selecting a subpopulation with extreme values.

An intuitive example of RTM in a non-academic setting was provided by the television show, *The Great British Bake Off*. Contestants on the show baked competitively and each week the worst competitor was eliminated while the best competitor was honored as the Star Baker. Understandably, in addition to the incredible skill required to bake at that level, there were also some random fluctuations in performance, and this contributed to the weekly performance of the bakers. Several weeks into the competition, the judges noticed

that the previous Star Bakers had relatively disappointing performances the week after they were named Star Baker. The judges deemed this the "Star Baker Curse". As a viewer with rudimentary statistics knowledge and a lack of superstitions, I recognized that the curse was actually RTM. The bakers were named Star Baker when they were at their best, with the random fluctuations acting in their favor. This luck was not likely to remain the next week and therefore their performance dropped. Of course, RTM did not *cause* the bakers to perform worse but the combined individual reasons created a pattern designated as RTM. It was the selection of the highest performers, based on a noisy metric, that highlighted the phenomenon of RTM.

In an education context, observed test scores are flawed metrics of knowledge or ability; observed scores are comprised of a true score and a random component. This random component can be due to luck, unobserved variables, or other short-term influences. Students can get extreme observed scores, or scores that are far from their average or true score, as a result of this random component. However, it is unlikely for individual students to consistently perform with the same extreme scores, so extreme scores are usually followed by less extreme scores. Similarly, students who had extreme scores on the second exam will likely have had less extreme scores on the first exam. When students are sorted into score bins by the observed scores on a single exam, they can end up in a score bin that is different than their true score bin, pushed over the edge by the random component.

RTM is more or less pronounced depending on the distribution of scores

and the magnitude of the random fluctuations in the observed scores. For a flat distribution with small fluctuations, the unlucky and lucky students in each bin tend to even out, resulting in little RTM. The exceptions are at the ends of the distribution, where ceiling and floor effects (such as maximum or minimum scores) could limit the extent of the luck. However, even in a flat distribution, if the random component is large enough to land students into bins other than the neighboring bins, the RTM in the middle of the distribution no longer evens out. As an example, each individual in a population could be randomly given a number one through five with roughly equal numbers of individuals given each number. If all of the individuals given the number four were given another random number (one through five), more of them would be given smaller numbers and thus the group's average would regress toward the mean.

RTM is exaggerated with uneven numbers of individuals in each bin. For normal distributions, the luck balances out near the mean of the distribution, but RTM is evident just outside of the mean score. This is because the students in an above-average bin, for example, are more likely to have been modestly performing students who did exceptionally well than high performing students on an off-day simply because the former outnumber the latter. Similarly, lower performing score bin averages are likely to increase upon retesting. Here again, the effects of RTM are particularly pronounced in the tails of the score distribution with ceiling and floor effects. Each time the students are grouped by their observed scores, (un)lucky students will end up in the "wrong

bins", causing the average scores for the groups to regress toward the mean on the next exam.

It is important to clarify that RTM does not cause inaccuracies; in fact, the regressed observed scores are usually closer to the true scores. However, RTM can be misinterpreted as a change in ability or true score. Similarly, data can be sorted in such a way that the RTM is exaggerated, which can lead to inaccurate results. This is the case in the streamline plots.

RTM is present in computations of both trajectories and streamlines, because both methods sort the students into groups by their observed scores. Trajectories group students only once in the lowest grade. Most of the RTM occurs between the sorting grade and the next as the observed scores regress toward the true scores. Therefore, the RTM in the first segment of a trajectory causes the scores to settle closer to the average true score for the group. Streamlines and the corresponding arrow plots, on the other hand, regroup students in every grade. This results in considerable RTM during each grade transition. The streamlines string together each of these regressions, exaggerating the RTM and producing inaccurate results. By identifying the influence of RTM in the streamline plots, I have developed a new method to better predict the flow of student scores.

To better understand the significance of RTM, I use conditional expectation values for the exam scores, as in Nesselroade et al. (1980) [83]. I invoke classical test theory [39, 40] to establish the relationship between the observed score, the true score, and the random error score. If x_i are the raw scores for

exam i, then $x_i = t_i + e_i$ where t_i is the true raw score and e_i is the random component. I can also use z-scores, z_i, such that $z_i = (x_i - \mu)/\sigma_{x_i}$, where σ_{x_i} is the standard deviation of x_i and μ is the mean raw score or expected score, $E(x_i) = \mu$. For z-scores [1], $E(z_i) = 0$ and $\sigma_{z_i} = 1$ for all i.

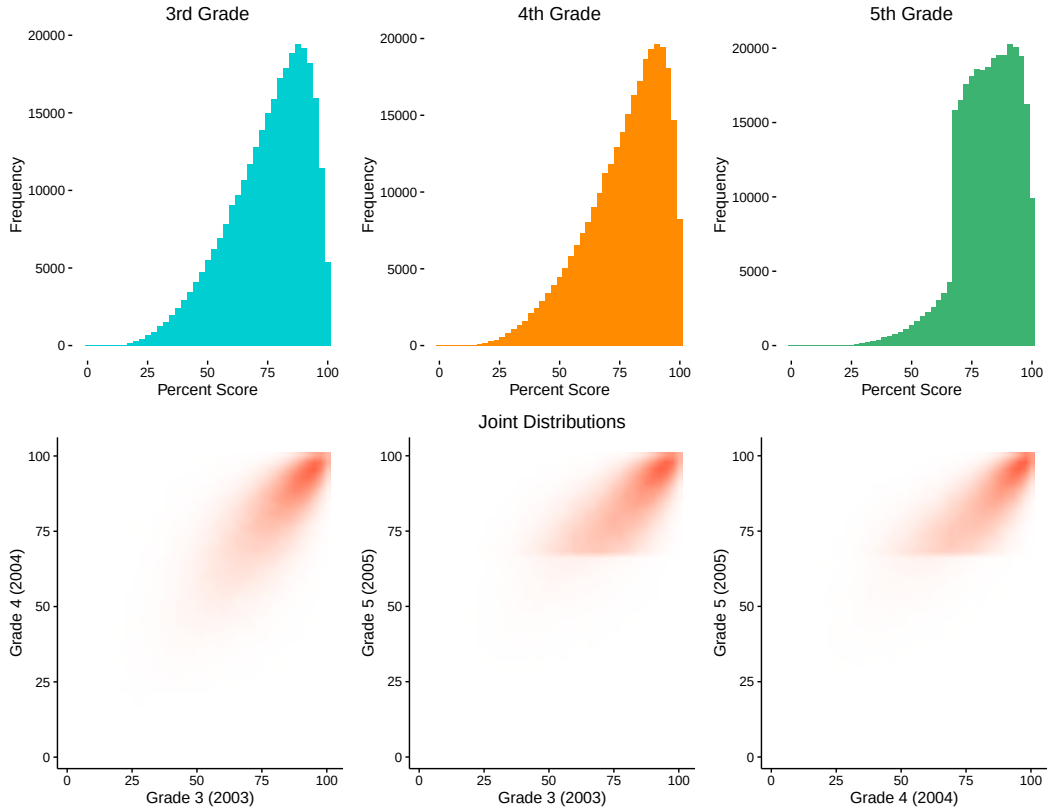

Figure 3.1: Distributions of mathematics TAKS percent scores for the cohort of students graduating in 2012 in 3rd, 4th, and 5th grade.

To make definite computations possible, I assume linear conditional expectation values. This assumption holds when the score distributions are

[1] Derivation in Appendix 1.

jointly normal[2], which is a common assumption [83, 84, 85, 86, 87], although it is not always valid for observed data. Nonetheless, a linear conditional expectation value is possible without a bivariate normal distribution. I investigated the normality and linearity of the raw mathematics TAKS scores for the cohort of 2012. For most of the exams, the distributions would have been roughly normal if not for the ceiling effects that cut off the higher end of the distribution. The distributions are therefore skewed. For 5th and 8th grade, the score distributions also have somewhat of a floor effect as a result of the Student Success Initiative. A small tail still exists below a cut-off score but the majority of the scores are squeezed above the cut-off, as seen in Figure 3.1. Despite these deviations from the normal distribution, the scores are fit no better by a higher order polynomial than by a linear regression. When comparing regressions of linear, quadratic, and cubic orders between two exams, the fit does not improve with additional terms and the coefficients for the higher order terms are small, as seen in Table 3.1. Therefore, the assumption of linear conditional expectation values provides an intuitive understanding of the theory without sacrificing too much accuracy. Working out the linear case provides a simple way to demonstrate why RTM causes problems in the streamlines and points the way to a solution that works in a more general setting. Select non-linear cases will also be addressed below.

Assuming a linear conditional expectation value for exam y given exam

[2]Derivation in Appendix 1.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
Intercept	-0.015	0.002	-7.80	6.32e-15
G3Score	0.747	0.002	357	<2e-16
G3Score2	0.010	0.002	5.45	5.02e-08
G3Score3	-0.007	0.001	-9.11	<2e-16

Table 3.1: Results from regression of the z-score in 4th grade with respect to the z-score in 3rd grade. Squared and cubic terms were significant but small. By adding quadratic and cubic terms, the R^2 value changed from 0.514 to 0.515.

x,

$$E(y|x) = \alpha + \beta x. \tag{3.1}$$

Both sides of Equation (3.1) can be multiplied by the probability density function, $P_x(x)dx$, and integrated by x to get

$$\begin{aligned} \int E(y|x)P_x(x)dx &= \int (\alpha + \beta x)P_x(x)dx \\ \int\int y\frac{P_{xy}(x,y)}{P_x(x)}P_x(x)dxdy &= \alpha + \beta \int xP_x(x)dx \\ E(y) &= \alpha + \beta E(x) \end{aligned} \tag{3.2}$$

Therefore, Equation (3.1) becomes

$$E(y|x) = E(y) - \beta E(x) + \beta x = E(y) + \beta(x - E(x)). \tag{3.3}$$

Multiplying both sides of Equation (3.1) instead by $xP_x(x)dx$ and integrating

over x gives

$$\int E(y|x)xP_x(x)dx = \int (\alpha + \beta x)xP_x(x)dx$$
$$\int\int y\frac{P_{xy}(x,y)}{P_x(x)}xP_x(x)dxdy = \alpha\int xP_x(x)dx + \beta\int x^2P_x(x)dx$$
$$E(xy) = \alpha E(x) + \beta E(x^2). \tag{3.4}$$

Equation (3.2) can be multiplied by $E(x)$ and then subtracted from Equation (3.4):

$$E(x)E(y) = \alpha E(x) + \beta E(x^2). \tag{3.5}$$

$$E(xy) - E(x)E(y) = \alpha E(x) + \beta E(x^2) - \alpha E(x) - \beta E(x)^2$$
$$E(xy) - E(x)E(y) = \beta(E(x^2) - E(x)^2).$$

The slope of the linear function, β, is then

$$\beta = \frac{E(xy) - E(x)E(y)}{E(x^2) - E(x)^2} = \frac{\sigma_{x,y}}{\sigma_x^2},$$

where $\sigma_{x,y}$ is the covariance of x and y, and σ_x^2 is the variance of x. Therefore, using the reliability ϱ or the Pearson correlation coefficient $\rho_{x,y}$, Equation (3.3) becomes

$$E(y|x) = E(y) + \frac{\sigma_{x,y}}{\sigma_x^2}(x - E(x)) \tag{3.6}$$
$$= E(y) + \varrho(x - E(x)) \tag{3.7}$$
$$= E(y) + \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x - E(x)). \tag{3.8}$$

In the case where x and y have the same true scores and expected scores μ, the expected change in score, or the *regression effect* [88], is given by:

$$E(y - x|x) = -(1 - \varrho)(x - \mu). \tag{3.9}$$

Therefore, if $x > \mu$ then y is expected to be less than x but greater than μ and if $x < \mu$ then y is expected to be greater than x but less than μ. So y is expected to be closer to the mean, but not go past it; y regresses toward the mean.

The relationship between an expected score and an observed score on a previous exam mirrors the relationship between true and observed scores, which is called *Kelley's equation* [89]:

$$\hat{T} = (1 - \varrho)\bar{X} + \varrho X. \tag{3.10}$$

where $\bar{X}$ is the average score, X is the observed score, and $\hat{T}$ is the expected true score.

The conditional expectation value expression in Equation 3.8 simplifies considerably with the substitution of z-scores. As noted above, z-scores have expectation values of zero and standard deviations of one, so the conditional expectation value of exam z_j given the score a on exam z_i is:

$$E(z_j|z_i = a) = \rho_{z_i,z_j} a \tag{3.11}$$

From the Cauchy-Schwarz inequality[3], $|\rho_{z_i,z_j}| \leq 1$; therefore, the score on exam j must be closer to the mean score, zero. Again, this is RTM. In particular,

the correlation coefficient between the exams determines the amount that the scores regress toward the mean. This can be extended to more than two years in several ways.

To model the trajectory plots using this method, I keep the selection criteria consistent throughout, since the groups of students are selected only by their initial score. The necessary conditional expectation values are:

$$E(z_2|z_1 = a) = \rho_{z_1,z_2} a$$

$$E(z_3|z_1 = a) = \rho_{z_1,z_3} a$$

$$E(z_4|z_1 = a) = \rho_{z_1,z_4} a$$

$$...$$

Stringing these scores together creates a sequence of anticipated scores, as seen in Table 3.2. If we assume that all of the exams have the same correlation coefficient $\rho_{z_1,z_j} = \rho$, this sequence would become $a, \rho a, \rho a, \rho a$, etc. Therefore the RTM takes place between the first two exams but not thereafter. The assumption of constant correlation coefficients is not valid for the data, but it helps to understand the basic structure of RTM.

Streamlines are constructed assuming a first-order Markov process, where the score in one year depends only on the score the year before. In this case we need to calculate the following conditional expectation values:

[3]Derivation in Appendix 1.

$$E(z_2|z_1 = a) = \rho_{z_1,z_2} a$$

$$E(z_3|z_2 = b) = \rho_{z_2,z_3} b$$

$$E(z_4|z_3 = c) = \rho_{z_3,z_4} c$$

$$...$$

To create a continuous streamline, the score for the one exam is used as the initial condition for the next, so that the expected values for the scores is

$$\varsigma_1 = a$$

$$\varsigma_2 = \rho_{z_1,z_2} a$$

$$\varsigma_3 = \rho_{z_2,z_3} \varsigma_2 = \rho_{z_2,z_3} \rho_{z_1,z_2} a$$

$$\varsigma_4 = \rho_{z_3,z_4} \varsigma_3 = \rho_{z_2,z_3} \rho_{z_2,z_3} \rho_{z_1,z_2} a$$

$$...$$

For streamlines, therefore, the scores are proportional to the product of previous correlation coefficients. For equivalent correlation coefficients ρ, the sequence becomes $a, \rho a, \rho^2 a, \rho^3 a$, etc. It is obvious in this case, because the absolute values of the correlation coefficients are less than one, that the scores continually regress towards the mean of zero. Table 3.2 compares the expressions in the sequences for streamlines and trajectories, and the sequences of scores using the calculated correlation coefficients from the data are shown in Table 3.5 (also compared to alternatively binned streamlines). Note that the linear correlation coefficients between exams in successive years are on the

Trajectories		Streamlines	
Expression	Equivalent Correlation Coefficients	Expression	Equivalent Correlation Coefficients
$\varsigma_3 = a$	a	$\varsigma_3 = a$	a
$\varsigma_4 = \rho_{3,4}a$	ρa	$\varsigma_4 = \rho_{3,4}a$	ρa
$\varsigma_5 = \rho_{3,5}a$	ρa	$\varsigma_5 = \rho_{4,5}\rho_{3,4}a$	$\rho^2 a$
$\varsigma_6 = \rho_{3,6}a$	ρa	$\varsigma_6 = \rho_{5,6}\rho_{4,5}\rho_{3,4}a$	$\rho^3 a$
$\varsigma_7 = \rho_{3,7}a$	ρa	$\varsigma_7 = \rho_{6,7}\rho_{5,6}\rho_{4,5}\rho_{3,4}a$	$\rho^4 a$

Table 3.2: Anticipated sequences of scores for trajectories and streamlines with respect to the initial score.

order of 0.7. Therefore within four years the product has dropped to around 0.25 of the initial score.

This simple linear theory demonstrates the quantitative differences in RTM in trajectory and streamline plots. This theory is very common in the RTM literature. Initially, I used a re-binning matrix to attempt to correct for RTM. Ultimately, I developed an alternative binning technique that reduced RTM, leading to the creation of alternatively binned (AB) streamlines.

3.2 Regression to the Mean in the Literature

RTM is known to be ubiquitous in many statistical fields, and many publications are dedicated to understanding and propagating the idea of RTM. The journal *Statistical Methods in Medical Research* had an issue exclusively focusing on RTM in medicine (Volume 6, Issue 2, April 1997). Several short papers have been written for the sole purpose of re-stating the widespread influence of RTM [90, 87]. To make the idea really stick, Rousseeuw targeted

researchers in general by acknowledging the influence of RTM on the error-prone paper selection process by editors, resulting in a worse selection of papers for publication than intended. Of course, the author had to excuse the journal that published his paper by saying, "a special exception should...be made for the exemplary periodical you are now consulting—that is, in which this article happened to land—which has an absolutely impeccable editorial record" [91].

Despite the numerous publications about RTM, the practical and theoretical understanding of RTM seems to be limited. A famous example of misinterpreting RTM was in Horace Secrist's book, *The Triumph of Mediocrity in Business* [92], in which he claimed that the data demonstrated the tendency for firms to become more mediocre over time, despite citing and knowing the work of Francis Galton, the discoverer of RTM [84]. Even within the community of scientists who write about RTM, the popular theory assumes bivariate normality (and therefore a linear conditional expectation value) and thus it does not apply to many realistic data distributions [93]. In general, RTM would occur in cases where

$$|E(Z_y|Z_x = z_x)| \leq |z_x|, \text{ and} \tag{3.12}$$

$$|z_x| - |E(Z_y|Z_x = z_x)| \text{ is an increasing function of } |z_x|, \tag{3.13}$$

but most papers do not prove these conditions are true for the data and only assume that they are [93]. Analytical understanding of RTM for non-normal cases is limited due to its complexity, but some studies of non-normal RTM do exist [94].

3.3 Matrix Re-binning

My first attempt at correcting for RTM followed from the idea that RTM was a direct result of students' observational score bins differing from their true score bins. Students were ending up in the "wrong bins", so if they could be re-binned into the correct bin then the RTM should be reduced. The main issue of course, is that the true scores are unknown.

Using several years of longitudinal testing data before the binning process would help to establish a good estimate for the true scores. Yet, the appeal of trajectories and streamlines is that they do not require several years of data to bin the students. Therefore, the re-binning process had to be designed to use only a year or two of data. Furthermore, it was not realistic to re-bin individual students when dealing with the student population for the entire state of Texas, so the re-binning technique needed to be performed on a large scale.

The two points to be addressed were (1) how many students should be moved from each bin and (2) which students should be moved. First, I decided that the students would be moved only to neighboring bins and the number moved would be dependent on the size of the score bin. I implemented the

following re-binning matrix:

$$M = \begin{pmatrix} 1+\epsilon & -\epsilon & & & & & \\ -\epsilon & 1+2\epsilon & -\epsilon & & & 0 & \\ & -\epsilon & 1+2\epsilon & -\epsilon & & & \\ & & & ... & & & \\ & & & -\epsilon & 1+2\epsilon & -\epsilon & \\ & 0 & & & -\epsilon & 1+2\epsilon & -\epsilon \\ & & & & & -\epsilon & 1+\epsilon \end{pmatrix} \tag{3.14}$$

which was multiplied by the vector of observational score bin cardinalities. This gave a vector of the new cardinalities after re-binning of:

$$N' = \begin{pmatrix} N_1 - (N_2 - N_1)\epsilon \\ N_2 - (N_1 + N_3 - 2N_2)\epsilon \\ N_3 - (N_2 + N_4 - 2N_3)\epsilon \\ ... \\ N_9 - (N_8 + N_{10} - 2N_9)\epsilon \\ N_{10} - (N_9 - N_{10})\epsilon \end{pmatrix}. \tag{3.15}$$

This re-binning matrix assumed that $2\epsilon N$ students in a middle bin had been wrongly binned and that ϵN students spilled into each neighboring bin. Therefore, these students needed to be moved back into the proper bin. Assuming the distribution of observed scores was roughly normal, this re-binning shrunk the width of the observed score distribution, with fewer students on the ends of the distribution. Essentially, this moved students closer to the mean so that hopefully they did not then regress toward the mean. The proportion of students from each bin that was moved, ϵ, was set to be a constant number with respect to score bin, although this did not need to be the case. Additionally, ϵ was limited by the differences in cardinalities of neighboring bins because the bins could not contain fewer than zero students. In particular, $\epsilon \leq N_{i\pm1}/N_i$. This was quite limiting, especially if this needed to hold for every score bin (in

the case of a constant ϵ). Ideally, ϵ would be a function of the standard error of measurement.

Next, the process of determining which students to re-bin needed to be decided. One option was to move the students randomly. This was unlikely to select the wrongly binned students. Another option was to move the students that were closest to the bin borders. This could be automated by sorting the students by their scores within each bin, indexing the students, and then moving the students with the appropriate index numbers. This was also unlikely to properly select students for re-binning. Students could have been sorted by the scores in the previous or next year and then indexed and moved. This incorporated a second year of data into the sorting process; a single exam was used for the initial binning and a second exam was used to decide which students to re-bin. While this method could have slightly improved the selection process, it was also unlikely to produce reliable results.

Each of these methods were tested with varying ϵ values to observe the resulting influence on the RTM. Figure 3.2 shows the cohort streamlines after re-binning the closest 10% of students to the bin borders when sorted by the next grade's score. The re-binning technique did little to reduce the RTM in most bins. In general, this technique was unable to produce reliable or accurate results. The technique was also fairly complex and difficult to automate. Ultimately, I found a different method to address RTM, which I called AB streamlines.

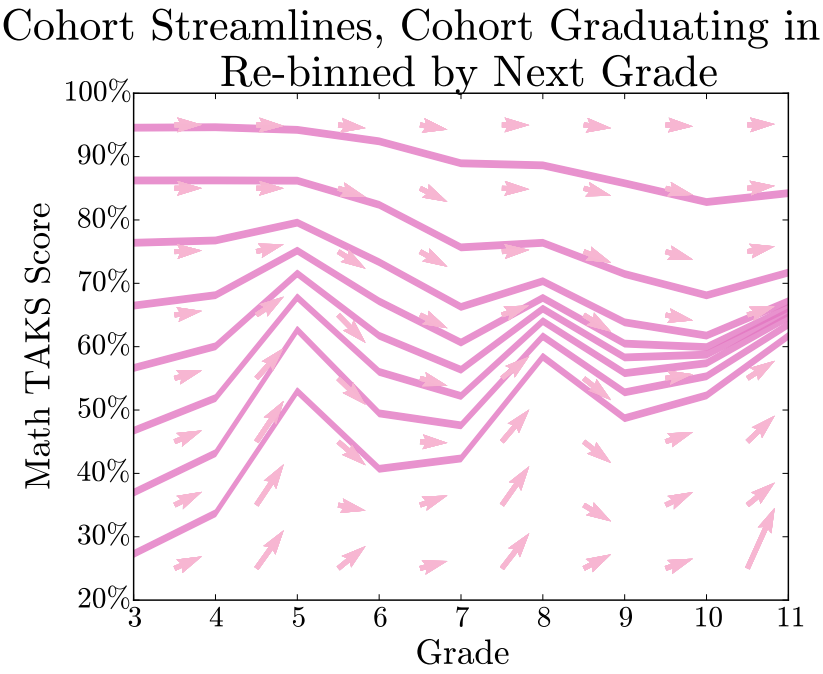

Figure 3.2: Re-binned cohort streamlines for the 2012 cohort. When sorted by the next grade's score, the closest 10% of students to the bin borders were moved to the neighboring bins.

3.4 Alternatively Binned (AB) Streamlines

The sequences of scores depicted in Table 3.2, particularly the sequence for trajectories, reveal that the most significant regression toward the mean takes place between the binning exam and the next exam. Therefore, by starting the analysis with the exam *after* the exam used for binning the students, the majority of the RTM effects will have resolved; this idea is the core of the alternative binning (AB) method. This modification is simple for trajectory plots: sort the students by the 3rd grade exam and plot the average scores between 4th and 11th grade. Applying the AB method to streamline plots is

slightly more involved because there are several exams used for binning.

For the streamlines described in the previous chapter, students are sorted by their scores on exam i for $i \in [3, 10]$ and the velocity, or average change in score, is calculated for those groups between exams i to $i + 1$. The scores from the binning exam are included in the velocity calculation. With alternative binning we delay the velocity calculation. In AB streamlines, students are sorted by their scores on exam i for $i \in [3, 9]$ and the velocities are calculated for those groups between exams $i + 1$ to $i + 2$. Excluding the binning exam scores from the velocity calculation reduces the magnitude of RTM. After the arrow plot is established from the velocity calculations, continuous streamlines are constructed.

AB streamlines can use the cohort or snapshot timings. AB cohort streamlines follow the same cohort of students throughout school, with slight fluctuations in the total number of students due to non-traditional students not advancing in grade sequentially; students are only required to have three consecutive years of data to be included for a segment of the plot and they can join the cohort at any time. An example of an AB cohort streamline for the cohort of 2012 can be seen in Figure 3.3. Students were sorted by their 3rd grade scores in 2003, and the velocities for those groups were calculated from 2004-2005. Then the students were sorted by their 4th grade score in 2004, and the velocities were calculated from 2005-2006, and so on.

AB snapshot streamlines capture the changing performance of a synthetic cohort of students within a three year window. The first year is used to

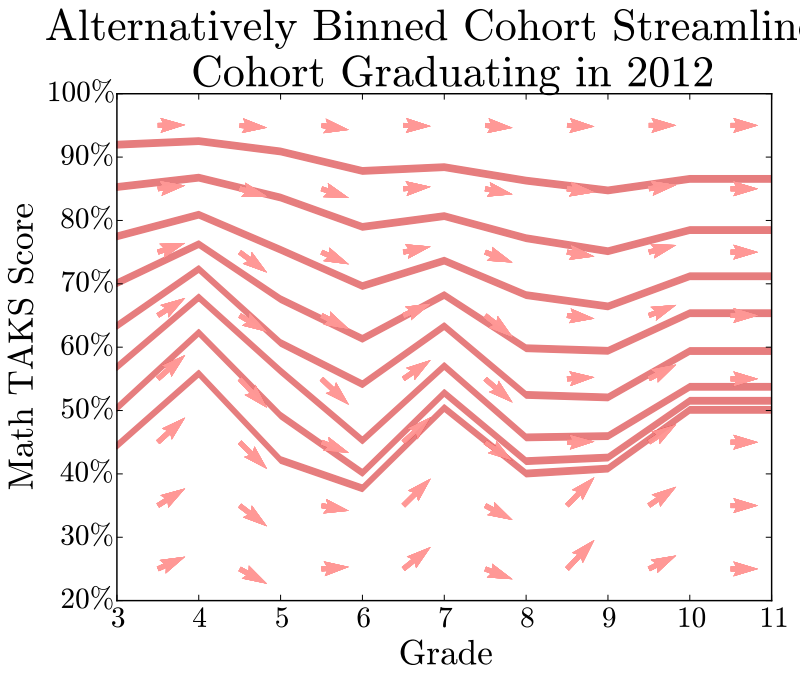

Figure 3.3: AB cohort streamlines for the cohort of 2012. The AB process reduces regression to the mean so that the streamlines no longer converge.

sort the students in each grade into groups determined by their percent scores. The second and third years are used to calculate the changing average scores for each group of students. AB snapshot streamlines use an accelerated longitudinal design with seven consecutive cohorts over three years, covering nine grades. The grade-cohort table for the AB snapshot streamline of 2003-2005 can be seen in Table 3.3. The AB snapshot streamlines and arrow plot for 2003-2005 can be seen in Figure 3.4. Students within each grade were sorted by their scores in 2003, and the velocities were calculated for those groups between 2004-2005.

	Grade								
	3	4	5	6	7	8	9	10	11
Cohort 2013	2003-2004	2004-2005	2005-2006						
Cohort 2012		2003-2004	2004-2005	2005-2006					
Cohort 2011			2003-2004	2004-2005	2005-2006				
Cohort 2010				2003-2004	2004-2005	2005-2006			
Cohort 2009					2003-2004	2004-2005	2005-2006		
Cohort 2008						2003-2004	2004-2005	2005-2006	
Cohort 2007							2003-2004	2004-2005	2005-2006

Table 3.3: The grades, cohorts (by graduation year), and school years of the data used for the AB snapshot streamline of 2003-2005.

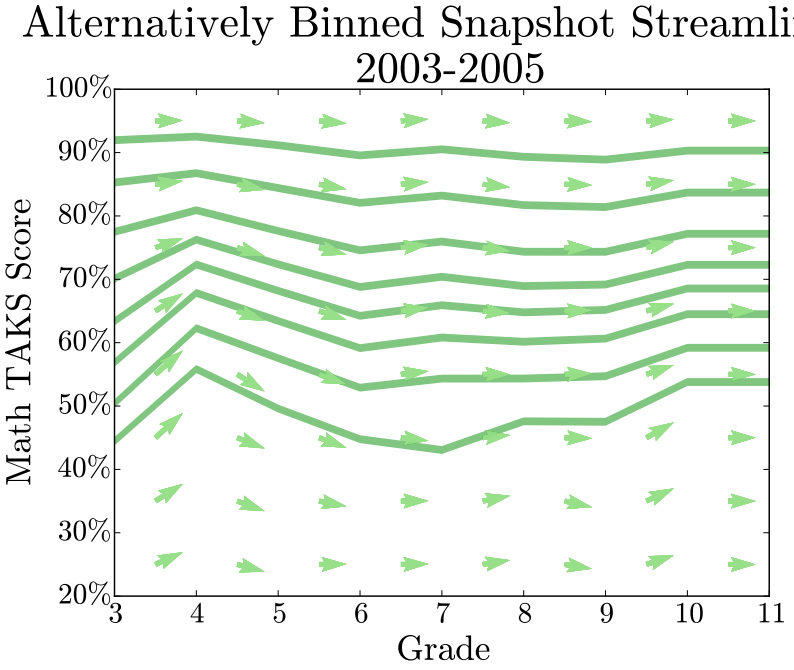

Figure 3.4: AB snapshot streamlines for the years between 2003 and 2005. The AB process reduces regression to the mean so that the streamlines no longer converge.

AB streamlines can be analyzed using the linear conditional expectation values, similar to the analysis of the trajectories and streamlines. For each sorting process, I calculate a pair of expectation values conditioned on the sorted scores. The necessary pairs of expectation values are:

$$\begin{cases} E(z_2|z_1 = a) = \rho_{z_1,z_2} a \\ E(z_3|z_1 = a) = \rho_{z_1,z_3} a \end{cases}$$

$$\begin{cases} E(z_3|z_2 = b) = \rho_{z_2,z_3} b \\ E(z_4|z_2 = b) = \rho_{z_2,z_4} b \end{cases}$$

$$\begin{cases} E(z_4|z_3 = c) = \rho_{z_3,z_4} c \\ E(z_5|z_3 = c) = \rho_{z_3,z_5} c \\ \cdots \end{cases}$$

The difference of the two scores in each pair is used as the slope for each of the vectors in the arrow plot of the AB streamline. To create the streamlines, each segment must be strung together by using the previous score as the initial condition for the next pair. Beginning with the third exam, there are two expectation values given for the same exam, one conditioned by the previous exam and one conditioned by the exam before that. To calculate the initial conditions (b, c, etc.), one simply has to set the two expectation values for the same exam equal to each other. For example, I would need to calculate what b would be such that $E(z_3|z_1 = a) = E(z_3|z_2 = b)$. This gives $b = \frac{\rho_{z_1,z_3}}{\rho_{z_2,z_3}} a$, which then allows me to calculate $E(z_4|z_2 = b)$. Using this method iteratively gives a sequence of scores in the AB streamlines of:

$$\varsigma_1 = a \quad \text{(Not included in the plot)}$$

$$\varsigma_2 = E(z_2|z_1 = a) = \rho_{z_1,z_2}a$$

$$\varsigma_3 = E(z_3|z_1 = a) = \rho_{z_1,z_3}a$$

$$\varsigma_4 = E(z_4|z_2 : E(z_3|z_2) = \varsigma_3) = E(z_4|z_2 : \rho_{z_2,z_3}z_2 = \rho_{z_1,z_3}a)$$
$$= E(z_4|z_2 = \frac{\rho_{z_1,z_3}}{\rho_{z_2,z_3}}a) = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_2,z_3}}a$$

$$\varsigma_5 = E(z_5|z_3 : E(z_4|z_3) = \varsigma_4) = E(z_5|z_3 : \rho_{z_3,z_4}z_3 = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_2,z_3}}a)$$
$$= E(z_5|z_3 = \frac{\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_3,z_4}\rho_{z_2,z_3}}a) = \frac{\rho_{z_3,z_5}\rho_{z_2,z_4}\rho_{z_1,z_3}}{\rho_{z_3,z_4}\rho_{z_2,z_3}}a$$

...

$$\varsigma_n = \frac{\rho_{z_{n-2},z_n}\rho_{z_{n-3},z_{n-1}}\rho_{z_{n-4},z_{n-2}}\cdots\rho_{z_1,z_3}}{\rho_{z_{n-2},z_{n-1}}\rho_{z_{n-3},z_{n-2}}\cdots\rho_{z_2,z_3}}a$$
$$= \frac{\prod_{p=1}^{n-2}(\rho_{p,p+2})}{\prod_{q=2}^{n-2}(\rho_{q,q+1})}a \quad (\forall n \geq 4).$$

In this sequence, the coefficients of the scores are ratios of Pearson correlation coefficients between the exams. Again, if I were to assume the same correlation coefficients between all of the exams, the values in the sequence would be $\rho a, \rho a, \rho a$, etc., which is identical to the sequence for the trajectory (excluding the initial exam).

Table 3.4 compares the trajectory, streamline, and AB streamline sequences for the special case where the correlation coefficients are the same between every exam. The trajectory and AB streamline sequences only regress

toward the mean between the 3rd and 4th grade exams, whereas the streamline sequence continuously regresses toward the mean. This is an indication that the AB streamlines may successfully reduce RTM to the natural amount. Table 3.5, shows the expressions and values for the anticipated scores in the trajectory, streamline, and AB streamline frameworks, using the observed Pearson correlation coefficients for the cohort of 2012. In the trajectory column, the RTM continues slightly each year, because the exams are not perfectly correlated, and they become less correlated over time. The most significant regression occurs between the 3rd grade and 4th grade exams. The streamline column shows that RTM effects are severe year after year. The scores are proportional to the *product* of the previous correlation coefficients. The AB streamlines mitigate this severe RTM because the previous score is multiplied by a *ratio* of similar correlation coefficients. The coefficient in the numerator is likely smaller, due to the extra year that has passed between the exams, so the scores still regress toward the mean; however, the rate of this regression is nearly identical to the rate in the trajectories.

The removal of the 3rd grade data (which is only used for sorting in the AB streamlines) from the sequences of scores in trajectories and AB streamlines greatly reduces the total amount of RTM throughout the grades, as it removes the extreme scores in the binning grade. As seen in Table 3.6, the total RTM from 3rd to 11th grade is 0.54 of the initial score for the 2012 trajectories and 0.53 of the initial score for the 2012 AB cohort streamlines. However, in relation to the 4th grade score, the 11th grade score is reduced

Scores for Equivalent Correlation Coefficients			
	Trajectories	Streamlines	AB Streamlines
ς_3	a	a	a
ς_4	ρa	ρa	ρa
ς_5	ρa	$\rho^2 a$	ρa
ς_6	ρa	$\rho^3 a$	ρa
ς_7	ρa	$\rho^4 a$	ρa
ς_8	ρa	$\rho^5 a$	ρa
ς_9	ρa	$\rho^6 a$	ρa
ς_{10}	ρa	$\rho^7 a$	ρa
ς_{11}	ρa	$\rho^8 a$	ρa

Table 3.4: Comparison between the score sequences within the trajectory, cohort streamline, and AB cohort streamline frameworks, assuming each pair of exams has the same correlation coefficient.

Trajectories		Streamlines		AB Streamlines	
Expression	Value	Expression	Value	Expression	Value
$\varsigma_3 = a$	a	$\varsigma_3 = a$	a	$\varsigma_3 = a$	a
$\varsigma_4 = \rho_{3,4}a$	$.73a$	$\varsigma_4 = \rho_{3,4}a$	$.73a$	$\varsigma_4 = \rho_{3,4}a$	$.73a$
$\varsigma_5 = \rho_{3,5}a$	$.67a$	$\varsigma_5 = \rho_{4,5}\varsigma_4$	$.53a$	$\varsigma_5 = \rho_{3,5}a$	$.67a$
$\varsigma_6 = \rho_{3,6}a$	$.65a$	$\varsigma_6 = \rho_{5,6}\varsigma_5$	$.38a$	$\varsigma_6 = \frac{\rho_{4,6}}{\rho_{4,5}}\varsigma_5$	$.64a$
$\varsigma_7 = \rho_{3,7}a$	$.63a$	$\varsigma_7 = \rho_{6,7}\varsigma_6$	$.30a$	$\varsigma_7 = \frac{\rho_{5,7}}{\rho_{5,6}}\varsigma_6$	$.62a$
$\varsigma_8 = \rho_{3,8}a$	$.62a$	$\varsigma_8 = \rho_{7,8}\varsigma_7$	$.23a$	$\varsigma_8 = \frac{\rho_{6,8}}{\rho_{6,7}}\varsigma_7$	$.59a$
$\varsigma_9 = \rho_{3,9}a$	$.58a$	$\varsigma_9 = \rho_{8,9}\varsigma_8$	$.18a$	$\varsigma_9 = \frac{\rho_{7,9}}{\rho_{7,8}}\varsigma_8$	$.57a$
$\varsigma_{10} = \rho_{3,10}a$	$.58a$	$\varsigma_{10} = \rho_{9,10}\varsigma_9$	$.15a$	$\varsigma_{10} = \frac{\rho_{8,10}}{\rho_{8,9}}\varsigma_9$	$.56a$
$\varsigma_{11} = \rho_{3,11}a$	$.54a$	$\varsigma_{11} = \rho_{10,11}\varsigma_{10}$	$.12a$	$\varsigma_{11} = \frac{\rho_{9,11}}{\rho_{9,10}}\varsigma_{10}$	$.53a$

Table 3.5: Comparison between the score sequences within the trajectory, cohort streamline, and AB cohort streamline frameworks, using the Pearson correlation coefficients computed from the data for the cohort of 2012.

Trajectories		AB Streamlines	
Compared with 3rd Grade Score	Compared with 4th Grade Score	Compared with 3rd Grade Score	Compared with 4th Grade Score
$\varsigma_3 = a$		$\varsigma_3 = a$	
$\varsigma_4 = .73a$	b	$\varsigma_4 = .73a$	b
$\varsigma_5 = .67a$	$.92b$	$\varsigma_5 = .67a$	$.92b$
$\varsigma_6 = .65a$	$.89b$	$\varsigma_6 = .64a$	$.88b$
$\varsigma_7 = .63a$	$.86b$	$\varsigma_7 = .62a$	$.85b$
$\varsigma_8 = .62a$	$.85b$	$\varsigma_8 = .59a$	$.81b$
$\varsigma_9 = .58a$	$.79b$	$\varsigma_9 = .57a$	$.78b$
$\varsigma_{10} = .58a$	$.79b$	$\varsigma_{10} = .56a$	$.77b$
$\varsigma_{11} = .54a$	$.74b$	$\varsigma_{11} = .53a$	$.73b$

Table 3.6: Sequences of anticipated z-scores for trajectories and AB cohort streamlines for the cohort of 2012. The first and third columns show the scores with respect to the 3rd grade score. The second and fourth columns show the scores with respect to the 4th grade score.

by only 0.74 and 0.73 of the initial score for the trajectories and AB streamlines, respectively. Therefore, the combination of alternative binning and the removal of the 3rd grade score from the plots greatly reduces the RTM in AB streamlines. The sequences for the trajectories and AB cohort streamlines are remarkably similar.

3.4.1 Non-Traditional Students

As with any longitudinal study, it is important to consider *attrition*, the decrease in the selected population over time. For the standardized testing data, attrition is usually a result of students moving to another state or more often, students veering from the *traditional* student pathway. I use the label traditional to indicate that the student progressed one grade each year and

therefore they stayed within a single cohort throughout their education. Non-traditional students are those that skip a grade or are retained in a grade, thereby moving to a different cohort. For the cohort of 2012, approximately 70% of the 3rd grade students remained in the cohort by 11th grade.

Due to the different binning processes in trajectory, streamline, and AB streamline plots, these techniques handle non-traditional students differently. In the standard full trajectories, students in a single cohort are sorted by their 3rd grade score, the first grade they take standardized exams. Therefore, students who join the cohort after 3rd grade, even if they join in 4th grade and are traditional thereafter, are not included in the analysis. If a student has a 3rd grade score but leaves the cohort in some later grade, they will be included for as long as they remain in the cohort. Trajectories do not need to be sorted by the 3rd grade score; they can be sorted by any grade. The trajectories for the cohort of 2012 sorted by their 9th grade scores can be seen in Figure 3.5. Scores for the grouped students are followed forward and backward in time. Regardless of the binning grade used, the students must have a score for the binning grade and will remain in the analysis for as long as they are members in the cohort. In addition, for the standard trajectories sorted by 3rd grade, I have required that the students also have 4th grade scores. This allows for the exclusion of the 3rd grade scores from the plot to reduce the RTM. In short, trajectory plots do not include non-traditional students in the analysis because they only use a single cohort of students and the students are only binned once.

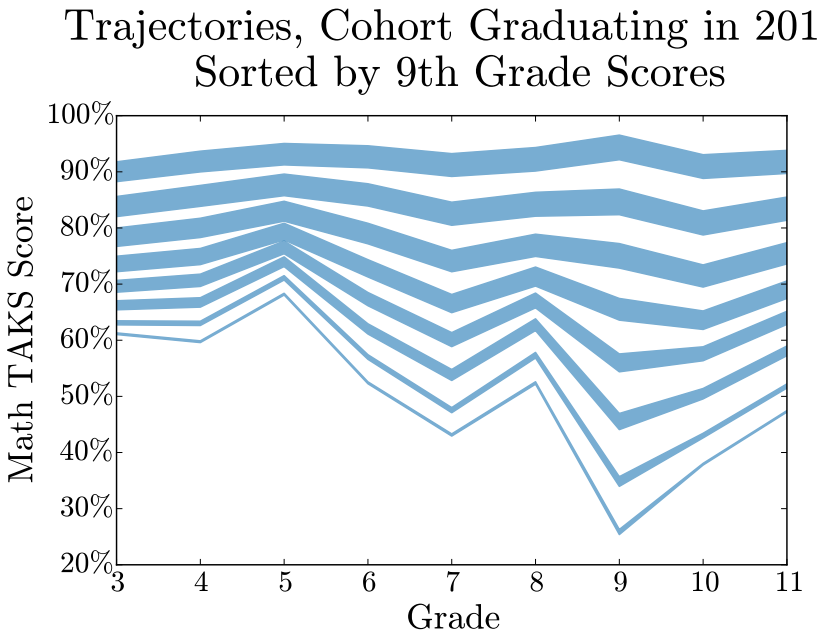

Figure 3.5: Trajectories for the cohort of students who graduated in 2012, sorted by their 9th grade scores. The scores exhibit regression to the mean in both directions away from the binning grade.

Streamline plots treat non-traditional students differently, depending on whether the plot is a cohort streamline plot or a snapshot streamline plot. Cohort streamline plots are very similar to trajectories in that they only use data from a single cohort. However, the binning process for streamlines is different than for trajectories and allows for the inclusion of some non-traditional students. Each segment of the plot, or grade transition, is calculated separately. For two consecutive grades, the students are sorted by the first grade and the velocity is calculated between the two grades. Therefore, if a student only had data for two consecutive grades, they would be included in for that segment of the plot. Therefore, students who are members in the cohort for

at least two consecutive grades will be partially included in the plot. Cohort streamlines are more inclusive than trajectories because cohort streamlines include temporary cohort members.

Snapshot streamline plots handle non-traditional students completely differently because instead of focusing on a single cohort, snapshot streamlines focus on a series of consecutive cohorts. Therefore, snapshot streamlines do not follow the traditional pathway. Students need to have scores in any two consecutive grades during the two years of interest to be included in that segment. Students are unlikely to be included for more than one segment in the plot because each segment represents a different cohort.

The alternative binning process changes the requirements for inclusion in the AB streamline plots. In both the cohort and snapshot AB streamlines, each segment requires *three* consecutive grades of data in a cohort, one for binning and two for calculating the velocity. Therefore, AB streamlines are less inclusive than streamlines but still more inclusive than trajectories.

The implications of the different requirements for inclusion of the three plotting schemes will be discussed further in the next chapter. The attrition of non-traditional students is likely to influence the results as these students are not a random sample. Furthermore, attrition can affect the interpretation of intervention effects. The Student Success Initiative directly causes failing students to repeat that grade and therefore creates non-traditional students. Attrition is therefore a consequence of the SSI that is captured in trajectories, streamlines, and AB streamlines differently.

3.5 Non-Linear Regression to the Mean

The linear theory shown above intuitively demonstrates the impact of RTM on trajectories and streamlines. In the interest of demonstrating the results and complexity of non-linear cases, I have derived the conditional expectation values for the quadratic case. I also reproduce the derivation of a more general case that assumes independent normally distributed random components as seen in the recently published paper by Schwarz and Reike (2017) [88].

3.5.1 Quadratic Conditional Expectation Value

Similar to the derivation above for the linear case, the process can be repeated for a conditional expectation with a quadratic term. Assuming a quadratic regression equation,

$$E(y|x) = A + Bx + Cx^2. \tag{3.16}$$

both sides can be multiplied by $P_x(x)dx$ and integrated:

$$\int E(y|x)P_x(x)dx = \int (A + Bx + Cx^2)P_x(x)dx$$

$$\int y\frac{P_{xy}(x,y)}{P_x(x)}P_x(x)dxdy = A\int P_x(x)dx + B\int xP_x(x)dx + C\int x^2P_x(x)dx$$

$$E(y) = A + BE(x) + CE(x^2). \tag{3.17}$$

Solving for A and substituting this into Equation (3.16) gives:

$$\begin{aligned} E(y|x) &= E(y) - BE(x) - CE(x^2) + Bx + Cx^2 \\ &= E(y) + B(x - E(x)) + C(x^2 - E(x^2)). \end{aligned} \tag{3.18}$$

Next, both sides of Equation (3.16) can be multiplied instead by $xP_x(x)dx$ and integrated:

$$\int E(y|x)xP_x(x)dx = \int (A + Bx + Cx^2)xP_x(x)dx$$

$$\int y\frac{P_{xy}(x,y)}{P_x(x)}xP_x(x)dxdy = A\int xP_x(x)dx + B\int x^2P_x(x)dx + C\int x^3P_x(x)dx$$

$$E(xy) = AE(x) + BE(x^2) + CE(x^3). \tag{3.19}$$

Equation (3.17) can be multiplied by $E(x)$:

$$E(x)E(y) = AE(x) + BE(x)^2 + CE(x^2)E(x). \tag{3.20}$$

Subtracting Equation (3.20) from Equation (3.19) and solving for B gives:

$$\begin{aligned} E(xy) - E(x)E(y) &= AE(x) + BE(x^2) + CE(x^3) \\ &\quad - AE(x) - BE(x)^2 - CE(x^2)E(x) \\ &= B(E(x^2) - E(x)^2) + C(E(x^3) - E(x^2)E(x)) \\ B &= \frac{E(xy) - E(x)E(y)}{E(x^2) - E(x)^2} - C\frac{(E(x^3) - E(x^2)E(x))}{E(x^2) - E(x)^2} \\ &= \frac{\sigma_{x,y}}{\sigma_x^2} - C\frac{\sigma_{x,x^2}}{\sigma_x^2}, \end{aligned}$$

where $\sigma_{x,y}$ is the covariance of x and y, and σ_x^2 is the variance of x. Substituting B back into Equation (3.18) gives:

$$E(y|x) = E(y) + \frac{\sigma_{x,y}}{\sigma_x^2}(x - E(x)) - C\frac{\sigma_{x,x^2}}{\sigma_x^2}(x - E(x)) + C(x^2 - E(x^2)). \tag{3.21}$$

Lastly, both sides of Equation (3.16) can be multiplied by $x^2 P_x(x)dx$ and integrated:

$$\int E(y|x)x^2 P_x(x)dx = \int (A + Bx + Cx^2)x^2 P_x(x)dx$$

$$\int y \frac{P_{xy}(x,y)}{P_x(x)} x^2 P_x(x)dxdy = A\int x^2 P_x(x)dx + B\int x^3 P_x(x)dx + C\int x^4 P_x(x)dx \tag{3.22}$$

$$E(x^2 y) = AE(x^2) + BE(x^3) + CE(x^4). \tag{3.23}$$

Equation (3.17) can be multiplied by $E(x^2)$:

$$E(x^2)E(y) = AE(x^2) + BE(x)E(x^2) + CE(x^2)^2. \tag{3.24}$$

Subtracting Equation (3.24) from Equation (3.23), substituting B, and solving for C gives:

$$\begin{aligned} E(x^2y) - E(x^2)E(y) &= AE(x^2) + BE(x^3) + CE(x^4) \\ &\quad - AE(x^2) - BE(x)E(x^2) - CE(x^2)^2 \\ &= B(E(x^3) - E(x)E(x^2)) + C(E(x^4) - E(x^2)^2) \\ \sigma_{x^2,y} &= B\sigma_{x,x^2} + C\sigma_{x^2}^2 \\ &= \frac{\sigma_{x,y}\sigma_{x,x^2}}{\sigma_x^2} - C\frac{\sigma_{x,x^2}^2}{\sigma_x^2} + C\sigma_{x^2}^2 \\ C &= \frac{\sigma_{x^2,y}\sigma_x^2 - \sigma_{x,y}\sigma_{x,x^2}}{\sigma_{x^2}^2\sigma_x^2 - \sigma_{x,x^2}^2}. \end{aligned}$$

Substitution into Equation (3.21) gives:

$$\begin{aligned} E(y|x) = E(y) &+ \frac{\sigma_{x,y}}{\sigma_x^2}(x - E(x)) - \frac{\sigma_{x^2,y}\sigma_x^2 - \sigma_{x,y}\sigma_{x,x^2}}{\sigma_{x^2}^2\sigma_x^2 - \sigma_{x,x^2}^2}\frac{\sigma_{x,x^2}}{\sigma_x^2}(x - E(x)) \\ &+ \frac{\sigma_{x^2,y}\sigma_x^2 - \sigma_{x,y}\sigma_{x,x^2}}{\sigma_{x^2}^2\sigma_x^2 - \sigma_{x,x^2}^2}(x^2 - E(x^2)). \end{aligned}$$

If x and y are jointly normal, then σ_{x,x^2} and $\sigma_{x^2,y}$ are zero and the expression simplifies to Equation 3.8.

3.5.2 Normally Distributed Random Components

A more general expression for the regression effect can be derived using Classical Test Theory while assuming independent normally distributed random components. Given the observed score $\mathbf{X}$, the true score $\mathbf{T}$ and the error term $\mathbf{E}$, for each measurement j,

$$\mathbf{X}_j = \mathbf{T} + \mathbf{E}_j.$$

The regression effect, for independent $\mathbf{E}_j$, is

$$\begin{aligned} R(x) &= E(\mathbf{X}_2 - \mathbf{X}_1 | \mathbf{X}_1 = x) = E(\mathbf{E}_2 - \mathbf{E}_1 | \mathbf{X}_1 = x) \\ &= -E(\mathbf{E}_1 | \mathbf{X}_1 = x) \end{aligned} \tag{3.25}$$

The conditional density of $\mathbf{E}_1$ given an observed score x is

$$\begin{aligned} P(\mathbf{E}_1 = e | \mathbf{T} + \mathbf{E}_1 = x) &= \frac{P(\mathbf{E}_1 = e) P(\mathbf{T} = x - e)}{P(\mathbf{T} + \mathbf{E}_1 = x)} \\ &= \frac{P_E(e) P_T(x - e)}{\int_{-\infty}^{\infty} P_E(e) P_T(x - e) de} \end{aligned} \tag{3.26}$$

where $P_E(e)$ is a normal density for the error term and P_T is the density of the true scores. Letting $t = x - e$, the regression effect is

$$R(x) = -\frac{\int_{-\infty}^{\infty} e P_E(e) P_T(x - e) de}{\int_{-\infty}^{\infty} P_E(e) P_T(x - e) de} = -\frac{\int_{-\infty}^{\infty} (x - t) P_E(x - t) P_T(t) dt}{\int_{-\infty}^{\infty} P_E(x - t) P_T(t) dt} \tag{3.27}$$

Note that the numerator is equal to the derivative of the denominator with respect to x multiplied by σ_E^2. Also, the denominator is equal to the density

of the observed scores, P_X. Therefore,

$$R(x) = \sigma_E^2 \frac{P_X'(x)}{P_X(x)} = \sigma_E^2 \frac{d}{dx} \ln[P_X(x)]. \tag{3.28}$$

Given $P_X(x)$, one could predict the score on a second measurement or the true score:

$$E(\mathbf{X}_2|\mathbf{X}_1 = x) = E(T|\mathbf{X}_1 = x) = x + R(x) = x + \sigma_E^2 \frac{d}{dx} \ln[P_X(x)]. \tag{3.29}$$

For the special case where P_X is a normal distribution, Equation 3.7 is recovered.

Chapter 4

Results

4.1 AB Cohort Streamlines

The accuracy of the Alternative Binning (AB) process can be determined by comparing the trajectories and AB cohort streamlines for the same cohort of students. It is important to note however, that attrition contributes to the differences between the results from trajectories and AB cohort streamlines.

Figures 4.1 and 4.2 show both the trajectories and AB cohort streamlines for the 2012 and 2013 cohorts, respectively. For both cohorts, the trajectories and AB cohort streamlines correspond extremely well. The persistent convergence that was an issue in the streamline plots is no longer evident in the AB streamlines. This is confirmation that the AB process reduces the RTM in the streamline plots to the amount seen in the trajectory plots, as estimated in the theoretical computations. The AB process accomplishes the task of producing accurate streamlines without excessive RTM.

In both cohorts, the AB cohort streamlines slightly underestimate the mathematics performance for the students after 5th grade, especially for the lowest performing score bins. Additionally, AB streamlines slightly overes-

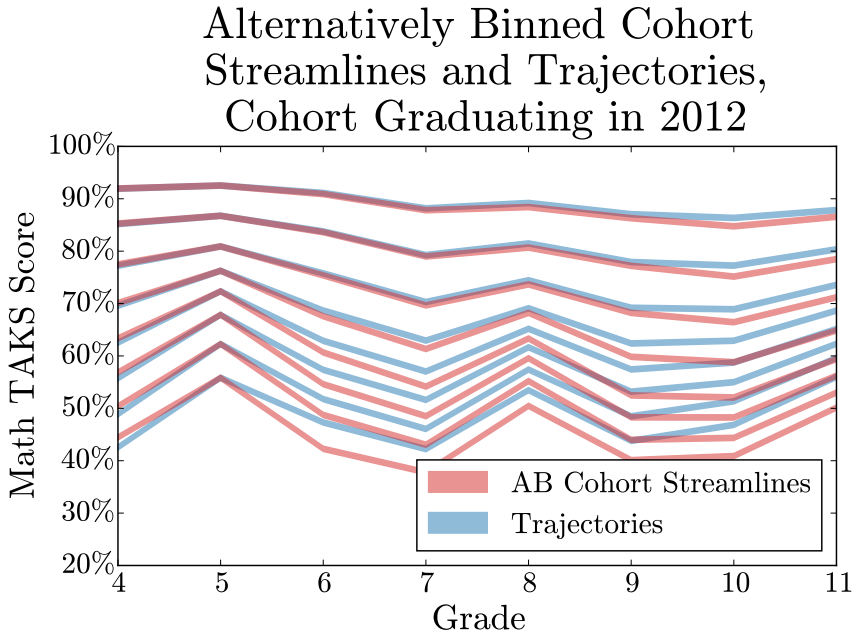

Figure 4.1: AB cohort streamlines and trajectories for the cohort of 2012. With the AB process, the cohort streamlines now reproduce the trajectories quite accurately.

timate the performance between 4th and 5th grade compared to the trajectories. For each grade and bin, I calculated the difference between the AB cohort streamline and trajectory and I also calculated the overall root mean square deviation (RMSD) between the two plotting methods. For the 2012 (2013) cohort, the maximum difference between the trajectories and AB cohort streamlines occurred for the 30-40% score bin in the 10th grade, with a difference in percent score of about 7.0 (7.5). The overall RMSD in percent score for both cohorts was only about 3.0. Considering that the score bins are separated initially by 10 percentage points, a variation of 3 percentage points (and often less) does not seem like a large variation.

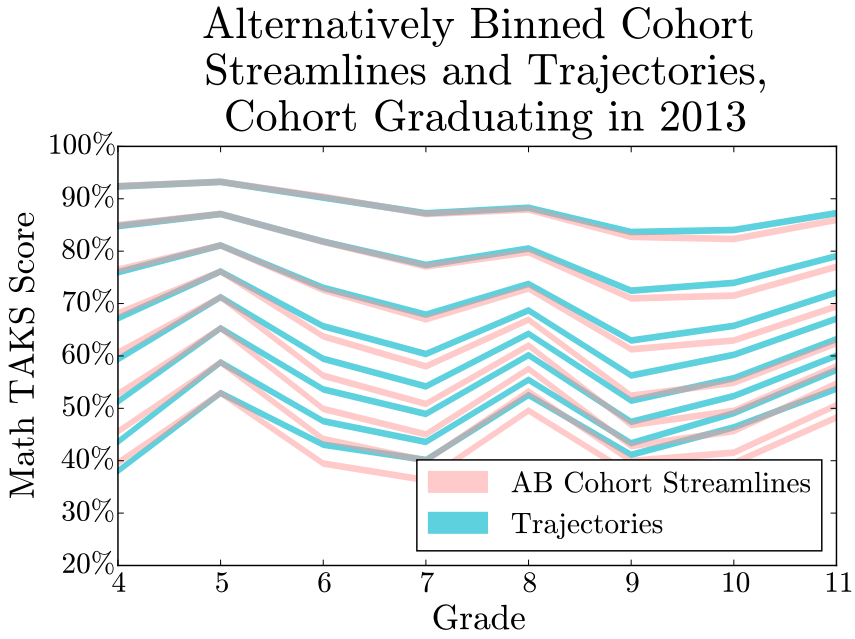

Figure 4.2: AB cohort streamlines and trajectories for the cohort of 2013. With the AB process, the cohort streamlines now reproduce the trajectories quite accurately.

This discrepancy between the AB cohort streamlines and the trajectories is likely a result of the differences in the inclusion of non-traditional students who leave or join the cohort. As discussed in the previous chapter, trajectories require the students to have 3rd and 4th grade scores whereas the AB cohort streamlines require students to have scores for grades $i-1$, i, and $i+1$ for the segment between grades i and $i+1$. Therefore, between 4th and 5th grade, the trajectories are slightly more inclusive than the AB cohort streamlines because the trajectories include students who have 3rd and 4th grade scores but no 5th grade scores. After 5th grade the AB cohort streamlines become more inclusive. For the segment between grades i and $i+1$,

students in the trajectories must have scores in grades 3, 4, i, and $i+1$ and students in the AB cohort streamlines must have scores in grades $i-1$, i, and $i+1$. This effectively means that the students in the trajectories must stay in the cohort from 3rd to $i+1$ grade, but the students in the AB cohort streamlines can join the cohort in grade $i-1$. Comparing the attrition for the different methods for the cohort of 2012 between 3rd and 9th grade, 80.8% of the students remained in the trajectories, 97.9% remained in the cohort streamlines, and 94.4% remained in the AB cohort streamlines. While AB cohort streamlines are slightly less inclusive than cohort streamlines, the AB process creates streamlines that are much more accurate.

The inclusion of non-traditional students seems to correlate with lower average performance, particularly for low performing students. The variation in the effect of this inclusion with score bin indicates that the population of non-traditional students is not random. This is expected since failing students may be required to repeat grades as mandated by the Student Success Initiative and mobility has been shown to correlate with lower mathematics performance [95]. AB cohort streamlines are more able to capture the performance of non-traditional students and the average scores for the lowest performing score bins are lower as a result.

4.1.1 Future Predictions

In 2012, the standardized exam program in Texas was changed to STAAR. With STAAR, the annual mathematics exams are no longer admin-

istered in high school; the students only take the STAAR mathematics exams in grades 3-8. Even so, not enough time has passed since STAAR began to create a full trajectory plot; the first cohort to take STAAR exams in 3rd grade completed 8th grade in 2017, so the data is not yet available. The AB snapshot streamlines provide a technique to predict scores with only three years of data, which is available. Figure 4.3 shows both the AB snapshot streamlines of 2012-2014 and the partial trajectories for the students who were 6th graders in 2015. The maximum difference in percent score between the predictions and the trajectories was about 13.0 in 6th grade, but in 4th and 5th grade was only 0.77. The RMSD in percent score was about 6.0 including 6th grade, only 0.3 without 6th grade.

As seen by the dip in the trajectories in 6th grade (2015), the scores for students in all score bins are considerably lower than predicted by the AB snapshot streamlines, even though the AB snapshot streamlines are using data from only one year earlier for that segment. This could suggest that the AB snapshot streamlines are not as accurate as hoped. However, an article published in *The Dallas Morning News* states that the STAAR mathematics exams in 2015 produced significantly lower scores compared to the previous year and that as a result, the Education Commissioner decided not to use the 2015 mathematics scores for accountability ratings or grade promotion [96]. It is unclear whether the lower performance in 2015 was due to an abnormally difficult exam or due to the effects of policy changes, such as HB-5. However, the test designers have been shown to produce consistent exams from year to

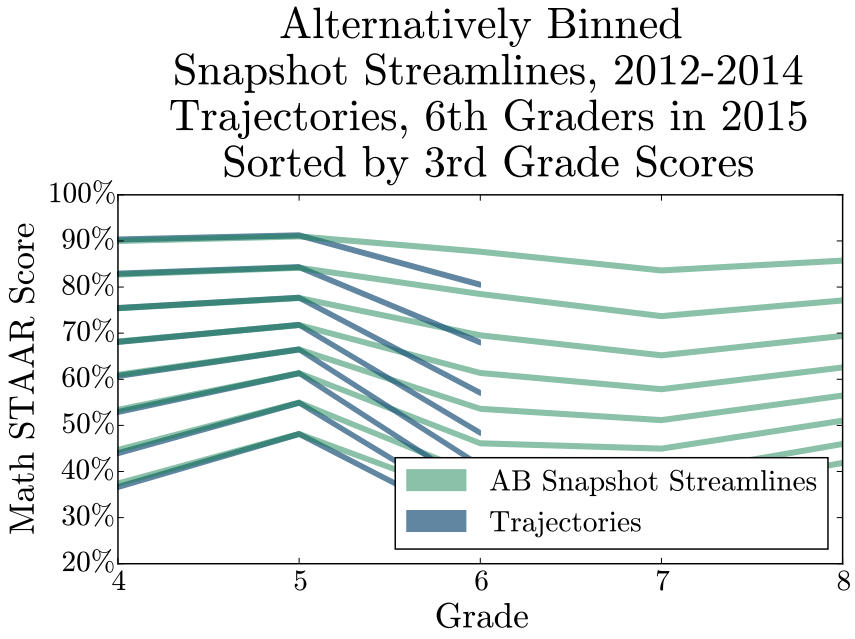

Figure 4.3: AB snapshot streamlines of 2012-2014 and the trajectories for the cohort of students that were 6th graders in 2015, using STAAR mathematics scores. The observed data for the 2015 6th graders is considerably lower than the predicted scores using the previous cohort.

year with TAKS and for the first few years of STAAR, so the chances that test design solely caused the drop in scores are small. Nonetheless, the documentation of this drop in performance provides an explanation for the differences between the trajectories and AB snapshot streamlines in 2015, and bolsters the predictive capabilities of the AB snapshot streamlines in the absence of period or cohort effects.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
Intercept	-0.078	0.002	-34.9	<2e-16
ReadScore	0.699	0.002	308	<2e-16
ReadScore2	0.088	0.002	37.7	<2e-16
ReadScore3	0.010	0.001	11.0	<2e-16

Table 4.1: Results from regression of the mathematics z-score in 4th grade with respect to the reading z-score in the same grade. By adding quadratic and cubic terms, the R^2 value changed from 0.4127 to 0.4199.

4.1.2 Sorting by Reading Score

In addition to using a previous mathematics exam to sort the students, the AB process could be accomplished by sorting the students by *any* exam that is reasonably well correlated (linearly, preferably) with the exams used to calculate the velocities. As demonstrated in Figure 4.4, the reading and mathematics exams have similar joint distributions to those between mathematics exams. Therefore, the reading score would be a good substitute for the previous mathematics score. This would reduce the number of years needed for each segment from three to two; using the first year's reading score to sort the students and then the first and second years' mathematics scores to compute the velocities. Table 4.1 shows the output from the regression of the mathematics score with respect to the reading score (and higher order terms).

The AB snapshot streamlines for 2008-2009, using the reading scores in each grade to sort the students into score bins, can be seen in Figure 4.5. The predictions do not have as similar a shape to the trajectories as did the AB streamlines that used the mathematics scores, but they do produce fairly accurate predictions in two years. The maximum difference in percent score

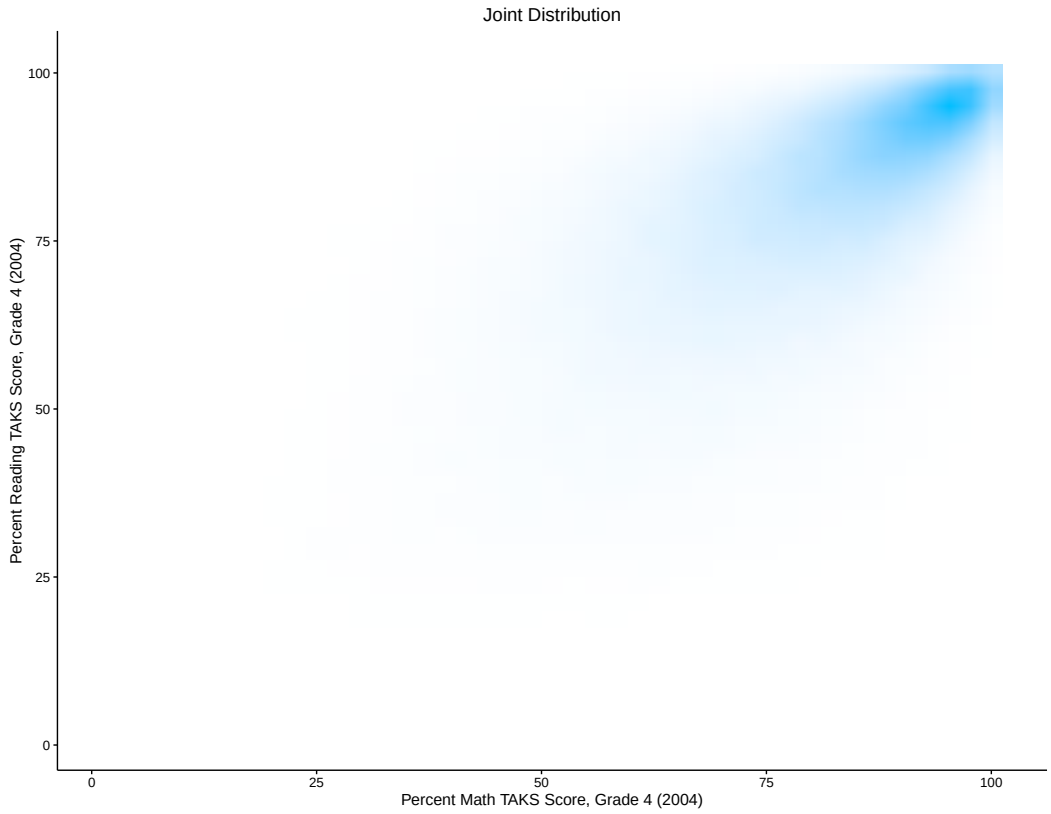

Figure 4.4: Comparison of score distributions for the mathematics and reading TAKS scores of the 4th graders in 2004.

and the RMSD for the full plots are actually smaller than for the AB cohort streamlines using the mathematics scores; the maximum difference in percent score was only 5.7 and the RMSD was only 2.2.

4.2 Student Success Initiative

AB cohort streamlines and AB snapshot streamlines are two tools that could be used to study the effects of the Student Success Initiative (SSI). AB

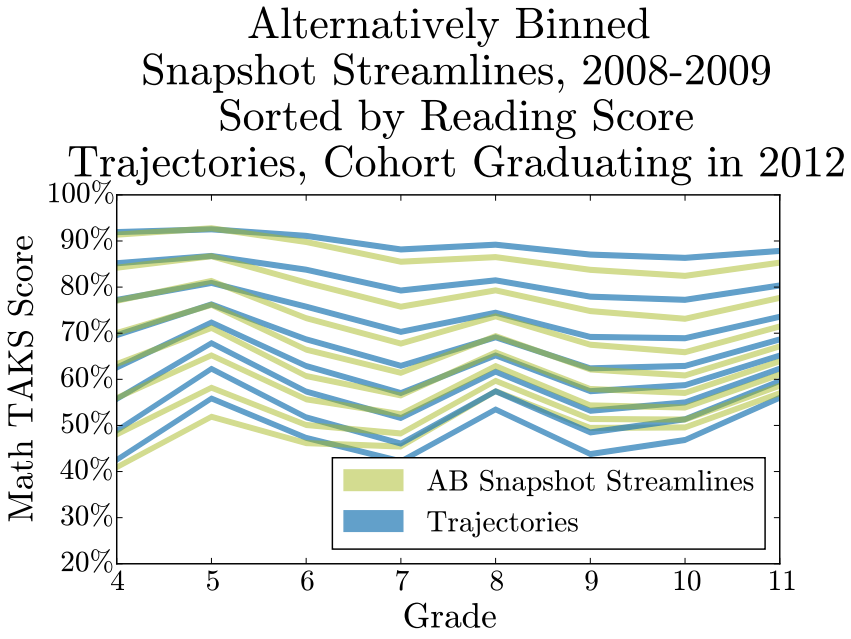

Figure 4.5: AB snapshot streamlines of 2008-2009 sorted by reading scores and trajectories for the cohort of 2012. By using the reading score for binning, accurate predictions can be made in only two years.

cohort streamlines are single cohort studies, which are useful for studying the evolution of a group of students over time. AB snapshot streamlines have an accelerated longitudinal design and can study a large age/grade range within a short period of time. I have investigated the consequences of the SSI using both of these methods.

Figure 4.6 shows the AB snapshot streamline for 2003-2005 along with the trajectory for the cohort of 2012. The SSI was implemented along with the cohort of 2012 and so it only reached K-5 students by 2005. Therefore, one could expect to see SSI effects for students throughout the 2012 cohort

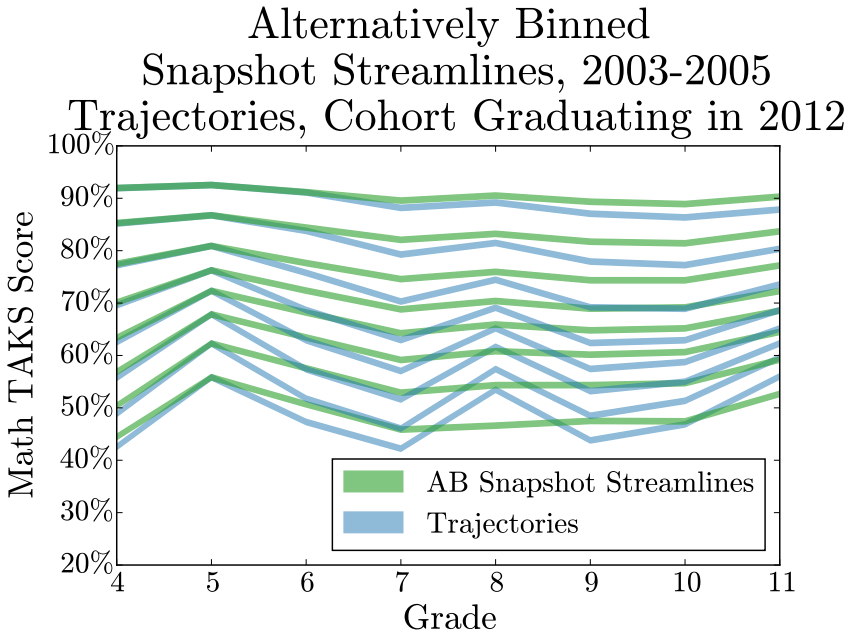

Figure 4.6: AB snapshot streamlines for 2003-2005 and trajectories for the 2012 cohort. The effects of SSI in 8th grade are captured in the trajectories but not the streamlines because the streamline data preceded SSI.

trajectory but only for the 3rd-5th grade students in the AB snapshot streamlines. In particular, the 8th grade peak seen in the trajectories but not the AB snapshot streamlines is expected since the 8th grade retention requirement and 8th grade Accelerated Math Instruction were implemented by 2008 but not in 2003-2005. This difference between the trajectories and AB snapshot streamlines demonstrates the effect of the SSI implementation; the trajectories, particularly the low performing students most affected by SSI, show a peak in performance in 8th grade (2008) whereas the 8th graders between 2003-2005 did not have improved performances.

If the 8th grade peak in the 2012 cohort trajectory is a result of the SSI, then the AB snapshot streamlines after 2008 should also show a peak in 8th grade. Figure 4.7 shows the AB snapshot streamline for 2007-2009 along with the same trajectory, for the cohort of 2012. These results do indeed show a peak in 8th grade for both the trajectories and the AB snapshot streamlines. This indicates that the 8th grade performance peak is likely due to the SSI. This analysis demonstrates that period or cohort effects can limit the accuracy of predictions made with only three years of data, although comparisons between AB snapshot streamlines of different years will identify the effects.

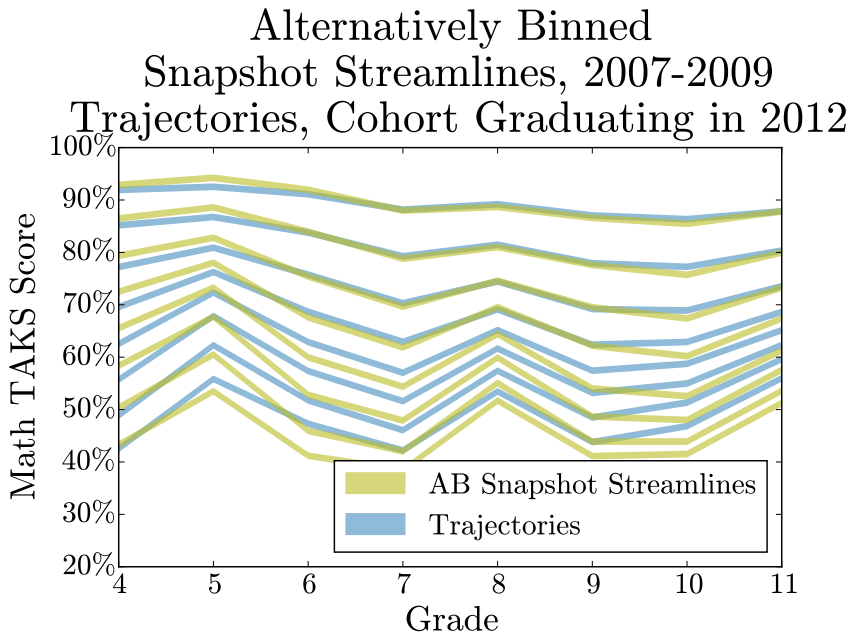

Figure 4.7: AB snapshot streamlines for 2007-2009 and trajectories for the 2012 cohort. Both methods capture the effects of SSI. The AB snapshot streamlines are able to predict the longitudinal data in only three years.

4.3 Demographic Differences

A subject of particular interest in education research, particularly STEM education research, is the study of differences in participation and performance for various demographic groups. The demographics are not thought of as the cause of these differences, rather the causes are various environmental influences associated with the demographic groups. These influences might be access to resources, family support, societal expectations, and so on. Nonetheless, differences in participation and performance appear in the data, highlighting unfortunate disparities in education.

The demographic categories that I chose to study were sex/gender (male or female, non-binary options were not available), socio-economic status (SES) (represented dichotomously by the free/reduced lunch status), and race/ethnicity (limited to White, Black/African American, Hispanic, or Asian). While SES is undeniably complicated in reality, for this analysis the SES variable was binary: zero if the student never received free or reduced lunch and one if the student ever received free or reduced lunch. These students were labeled not low-income and low-income, respectively. Therefore, this analysis oversimplifies the influence of SES. It is also important to note that sex/gender and race/ethnicity are also nuanced characteristics that are simplified in this analysis.

Table 4.2 shows the percentages of each demographic group during two periods for the same cohort; in both columns, the students were freshman in 2008-2009. The first column includes all of the first time 9th graders in

	Percent of Cohort	
Group	All 9th Graders	Traditional
Female	48.7	50.3
Male	51.3	49.7
Asian	3.7	4.1
Black	14.5	13.7
Hispanic	45.1	42.3
White	36.4	39.6
Not Low-Income	45.3	53.4
Low-Income	46.6	46.6

Table 4.2: Demographic breakdown of the 9th graders in 2008-2009. The traditional cohort limits the total to the population of students who remained in the cohort, graduating in 2012.

the 2008-2009 school year. Therefore, this includes the traditional cohort of students that graduated in 2012, as well as the students who began in that cohort but left the cohort during high school. The second column represents just the traditional cohort of students that graduated in 2012. Therefore, this includes students who were freshmen in 2008-2009, sophomores in 2009-2010, juniors in 2010-2011, and seniors in 2011-2012, comprising 74% of the total 9th grade population. The differences represent the students who left the cohort. The population that left the cohort was slightly more male, Hispanic or Black, and low-income. Some students in the total population had missing demographic data. Further analysis about participation in the context of physics courses will be discussed below.

Observed performance disparities can be studied using trajectory plots. The trajectories for the cohort of 2012, disaggregated by sex, SES, and race/ethnicity can be seen in Figures 4.8, 4.9, and 4.10. The male and female

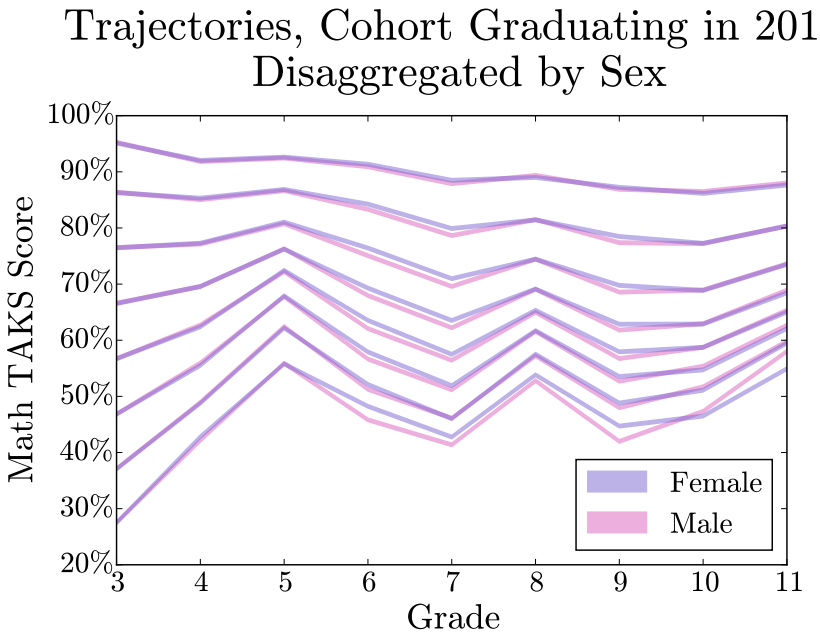

Figure 4.8: Trajectories for the cohort of 2012 disaggregated by sex. There are minimal performance disparities associated with differences in sex.

students had nearly equivalent performance over time (RMSD = 0.86), with men performing slightly worse at times, especially the low-performing middle-schoolers. SES appears to be correlated with differences in the scores, with low-income students consistently under-performing compared to not low-income students in the same initial score bin (RMSD = 3.5). Race/ethnicity disaggregation also shows differences in scores, although there may be combined effects between SES and race/ethnicity. Therefore, to separate these effects (keeping in mind the oversimplification of SES), Figures 4.11 and 4.12 show the trajectories for the low-income and not low-income groups, disaggregated by ethnicity. The racial/ethnic disparities become smaller when controlling

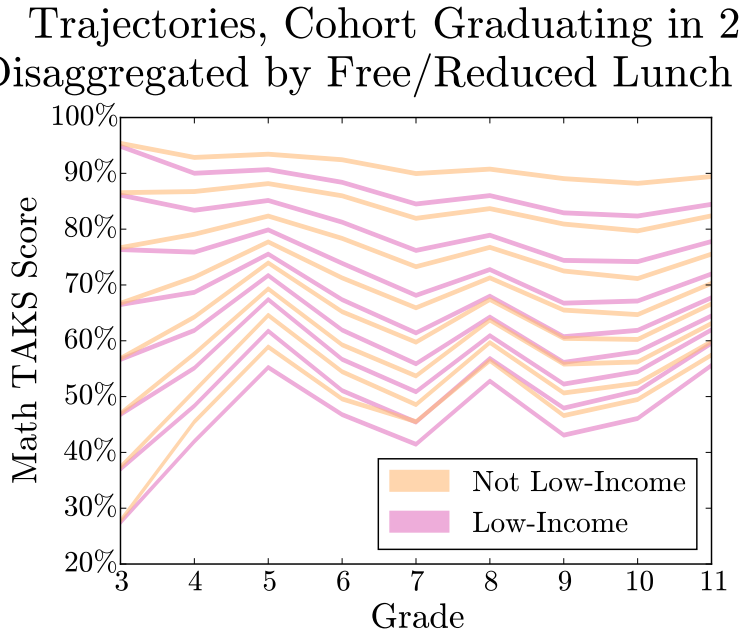

Figure 4.9: Trajectories for the cohort of 2012 disaggregated by SES. Students who received free or reduced lunch were considered low-income, otherwise students were considered not low-income. Not low-income students perform better than their low-income counterparts.

for SES, although Asian students still out-perform their peers. This should not be interpreted as an effect of race/ethnicity; rather there are likely latent factors that correlate with race/ethnicity and with varying performance levels.

4.4 Physics Course Taking

In addition to disaggregating students by demographic variables, students could be disaggregated by course-taking. While this would not prove causality between course taking and mathematics performance, it could establish a correlation. In particular, I was most interested in physics course

Group	% of Cohort	% of Group Taking...		
		Basic Physics	IPC	AP Physics
Female	50	76	36	4
Male	50	75	39	7
Asian	4	81	20	18
Black	14	69	48	2
Hispanic	42	76	40	4
White	40	77	32	6
Not Low-Income	53	78	31	7
Low-Income	47	72	45	3

Table 4.3: Course taking percentages for the traditional cohort of 2012.

taking. In the dataset, there are three main options for physics courses. Integrated physics and chemistry (IPC) is a low level science course, with the majority of enrollment occurring in 9th grade. Basic physics is the standard high school physics course, usually taken in 11th grade. Advanced placement (AP) physics is a college-level physics course, which is mostly attended by 12th graders. Students can take more than one of these options. The percentages of disaggregated students in the cohort of 2012 who took each course is shown in Table 4.3. Specifically the population was limited to students who were 9th graders in 2008-2009, 10th graders in 2009-2010, 11th graders in 2010-2011, and graduated in 2012. Approximately 37% of the cohort took IPC, 76% of the cohort took basic physics, and 5% of the cohort took AP physics.

Figure 4.13 shows the trajectories for the cohort of 2012 disaggregated by the highest level of physics course taking. The students in yellow took up to IPC, the students in green took up to basic physics, and the students in pink

took up to AP Physics. The students in each course taking group were sorted into score bins by their 3rd grade mathematics TAKS scores. Note that there are students that were eventually enrolled in each option within each score bin in 3rd grade (excluding the lowest three bins which had too few students for the analysis); some low performing students in 3rd grade took AP physics and some high performing students in 3rd grade took IPC. Despite students from each course performing similarly in 3rd grade (within a score bin), the longitudinal performance of the students in each course diverges. As expected, the AP students have the highest average mathematics scores with respect to score bin and the IPC students have the lowest scores with respect to score bin. The RMSD between the IPC and Basic Physics trajectories was 4.5, between the Basic Physics and AP Physics trajectories was 9.5, and between the IPC and AP Physics trajectories was 13.3.

It is obviously incorrect to state that physics course taking *causes* differences in mathematics score. After all, the variation in mathematics score occurs before the students enroll in any physics course. Instead, these differences are likely demonstrating the influences of a latent variable. This could be parental involvement, tutoring, mathematics interest, or some other unmeasured factor.

I used a logistic regression to study the probability of taking AP physics given demographic and course taking backgrounds. I tried several combinations of covariates; one example used the students' sex, ethnicity/race, SES, and prior course taking to determine the log odds of taking AP physics. If

$\pi_i = Pr(Y_i = 1|X_i = x_i)$ is the probability of a student taking AP Physics given the vector X of conditions, then:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_5 x_5 + \beta_6 x_6 \quad (4.1)$$

where x_1 is a binary sex variable (zero for male, one for female), x_2 is a binary SES variable (zero for not low-income, one for low-income), x_3 through x_5 are ethnic/racial dummy variables (all zero for White, each equaling one for Black, Hispanic, or Asian), and x_6 is a binary course taking variable. The courses I chose between in the model were Algebra I (in 9th grade), IPC, Geometry, Computer Science, AP Computer Science, AP Chemistry, and AP Biology. The odds ratios are the exponentiated coefficients.

Table 4.4 shows the odds ratios for taking AP physics given by the logistic regression. In summary, women and low-income students are half as likely to take AP as men and not low-income students, respectively. Asian students are three times as likely to take AP physics as White students. Black students are 2/3 as likely compared to White students. Hispanic students are roughly as likely to take AP physics as White students. The biggest contributor to taking AP physics according to the model is the previous course taking. If the student did not take Algebra I in 9th grade (meaning they likely took it in 8th grade), then they were seven times as likely to take AP physics compared to students who waited until 9th grade to take Algebra I. If students did not take IPC (and instead likely took basic physics or skipped straight to AP), they were almost twelve times as likely to take AP physics as IPC students.

Covariate	Alg I (9th Grade)	IPC
Female (vs. Male)	0.575	0.548
Low-Income (vs. Not)	0.514	0.502
Asian (vs. White)	3.021	3.394
Black (vs. White)	0.648	0.634
Hispanic (vs. White)	.995	0.898
No Course (vs. Yes)	6.915	11.68

Table 4.4: The odds ratios for taking AP physics, which equal the exponentiated coefficients from the logistic regression in Equation 4.1.

4.5 Course Requirements

Course taking patterns are likely to change as high school graduation course requirements change. This hypothesis can be tested twice within the ERC dataset, as course requirements in Texas changed in 2007 and 2014. Before 2014, three graduation plans were defined, the Minimum High School Program (MHSP), the Recommended High School Program (RHSP), and the Distinguished Achievement Program (DAP). The MHSP was designed for students who did not intend to pursue a higher degree and thus the MSHP was the minimum course load needed to graduate with a high school diploma. The RHSP was designed for students who were preparing for college. The DAP did not have different STEM course requirements than the RHSP but in addition, students had to complete four advanced measures, choosing between advised research projects, AP/IB courses, or the PSAT.

The course requirements between 2007-2014 were nicknamed 4x4 because students were required to take four courses in each of the four core

subjects: mathematics, science, English, and social studies. Students that pursued the RHSP or DAP were required to take Physics. The 4x4 corresponded with the prerequisites listed by the University of Texas at Austin [97] and other universities. While high school graduation requirements have moved away from the 4x4, college prerequisites have not.

House Bill 5 (HB-5), which was implemented in 2014, changed the structure of high school graduation requirements to the Foundation program. The Foundation program is designed to be a flexible program that allows students to pursue their interests. The universally required courses are limited but students choose one or more *endorsements* to supplement the Foundation (although students can graduate without an endorsement). The endorsement fields include STEM, business and industry, public services, arts and humanities, and multi-disciplinary. Schools are only required to offer the multi-disciplinary endorsement. All endorsements require students to attend a total of four mathematics courses and four science courses. Physics and Algebra II are no longer required specifically, although the TEA informs students that most colleges list Algebra II as a prerequisite. Students earn a Distinguished Level of Achievement if they complete the Foundation requirements, take four science and mathematics classes (including Algebra II), and complete one endorsement. The Distinguished Level of Achievement is necessary for students to qualify for admission to a Texas public university through the top 10 percent automatic admission law.

Table 4.5 shows the mathematics and science course requirements dur-

	Before 4x4 (2001-2006)		4x4(2007-2014)		HB-5 (2014-)
Field	MHSP	RHSP/DAP	MHSP	RHSP/DAP	Foundation
Math	Algebra I Geometry Elective	Algebra I Geometry Algebra II	Unchanged	Algebra I Geometry Algebra II Elective	Algebra I Geometry Elective
Science	Biology IPC	Biology 2/3 of IPC, Physics, & Chemistry	Unchanged	Biology Chemistry Physics Elective	Biology IPC Elective

Table 4.5: Mathematics and science course requirements in Texas between 2001 and the present [5].

ing the three periods since 2001. Algebra I, Geometry, and Biology have consistently been required for graduation. Algebra II was required by the RHSP until 2014. IPC, Physics, and Chemistry requirements were different for each graduation plan. The RHSP/DAP between 2007-2014 was the only plan that required Physics. For all other plans, students could instead take IPC (or IPC and Chemistry) to fulfill the science requirement.

I investigated physics course taking patterns from 2003 to 2016, as the enrollment for these courses was affected by the changing requirements. Figure 4.14 shows the total number of students who took AP Physics, Physics, or IPC each year. The demographic files in the ERC only went back to 2008, so I was unable to disaggregate by grade or demographics before then. The impact of the 4x4 requirements is evident by the steep decline in IPC enrollment and the growing physics enrollment. The early consequences of HB-5 can be seen by the reversal of these trends as well as an increase in AP Physics enrollment.

4.6 Teacher Certification

Teacher certification varies by program, type, duration, and field. There are numerous programs through which a teacher can become certified and these programs can be divided into two categories: university-based and alternative. NCLB mandated that teachers earn a bachelor's degree in addition to earning a certification. University-based programs allow future teachers to simultaneously earn their bachelor's degree and their certification. These student-teachers learn about the subject material as well as teaching practices and pedagogy. If a person interested in teaching already has a bachelor's degree, they can enroll in an alternative teacher certification program. Alternative programs provide a faster pathway to teacher certification, often utilizing a hands-on learning approach by assigning teachers to classrooms while they work toward the standard certification. While the teacher is still in the program, they are given a probationary certification.

Within the STEM fields there are several certification field options, focusing on one or more fields. The available STEM certifications are chemistry, computer science, life sciences, mathematics, and physical sciences, as well as composite certifications combining the above fields. Teachers of STEM courses do not necessarily have a STEM certification. Table 4.6 shows the percentage of active classroom teachers who taught the 2012 through 2015 cohorts who also had *any* STEM field certification. AP courses have the most qualified teachers, while the popular entry level courses have fewer qualified teachers, an issue that became worse over time. Only 80.7% of the teachers

Percentage of Teachers with *Any* STEM Certification by Student Cohort				
Subject	2012	2013	2014	2015
IPC	80.7	77.9	72.2	67.7
Chemistry	87.1	83.7	81.9	81.3
Computer Science	92.2	93.7	91.1	81.9
Physics	89.3	87.4	85.3	82.5
Geometry	88.5	88.3	87.6	86.4
Algebra I	88.4	88.3	88.3	86.9
Biology	88.4	87.8	87.6	87.1
Algebra II	92.1	92.3	91.1	89.4
Pre-Calculus	96.8	96.4	95.0	95.0
AP Computer Science	97.6	98.6	98.7	97.5
AP Biology	98.3	98.8	99.1	98.3
AP Physics	99.0	99.5	98.3	98.8
AP Chemistry	99.2	99.6	99.4	99.0
AP Calculus	99.0	98.2	98.6	99.1

Table 4.6: Percentages of teachers teaching the 2012-2015 cohorts who have a STEM certification. Lower level courses have a smaller proportion of STEM certified teachers.

who taught IPC to the cohort of 2012 were qualified to teach that subject, and this dropped to 67.7% for the cohort of 2015. Therefore, the students who were perhaps struggling the most with STEM also had a bigger chance of having an unqualified teacher.

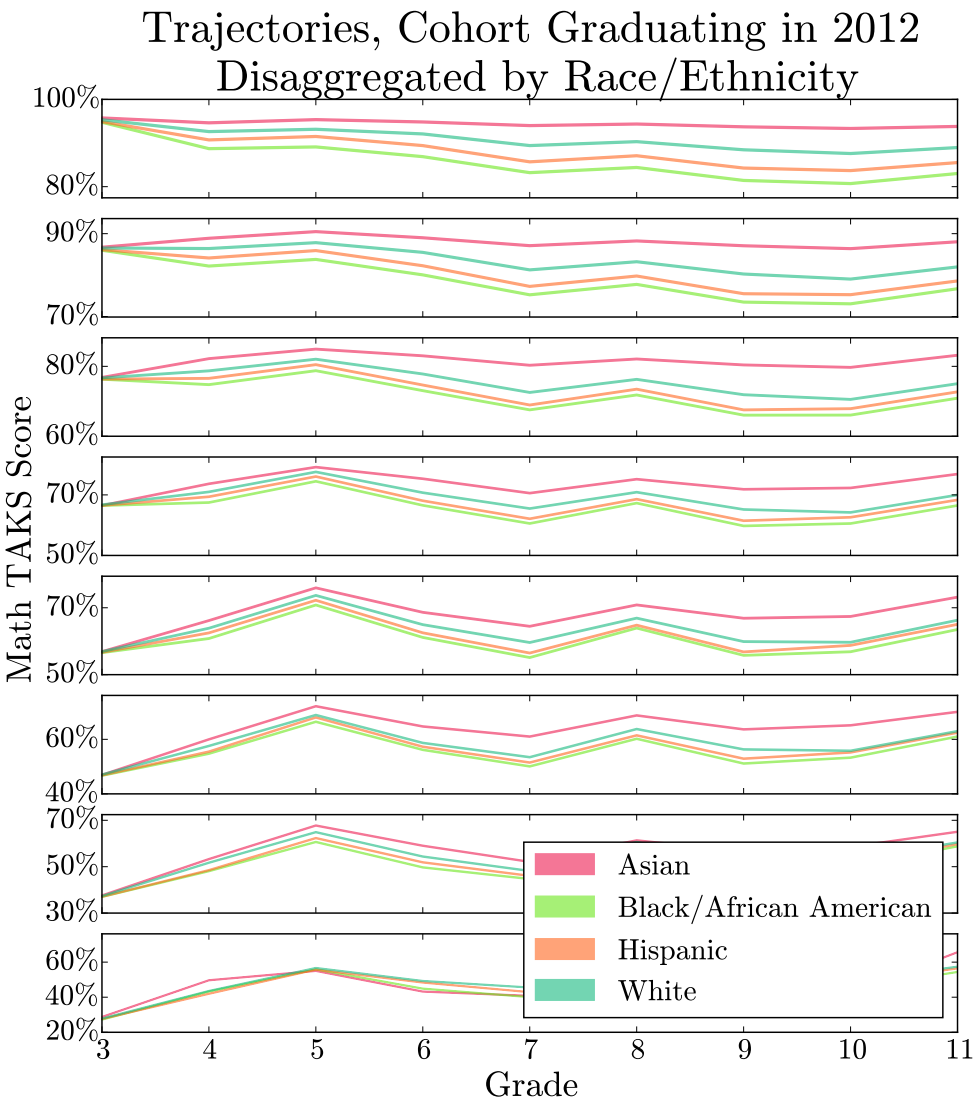

Figure 4.10: Trajectories for the cohort of 2012 disaggregated by race/ethnicity. There are performance disparities that may diminish within SES groups.

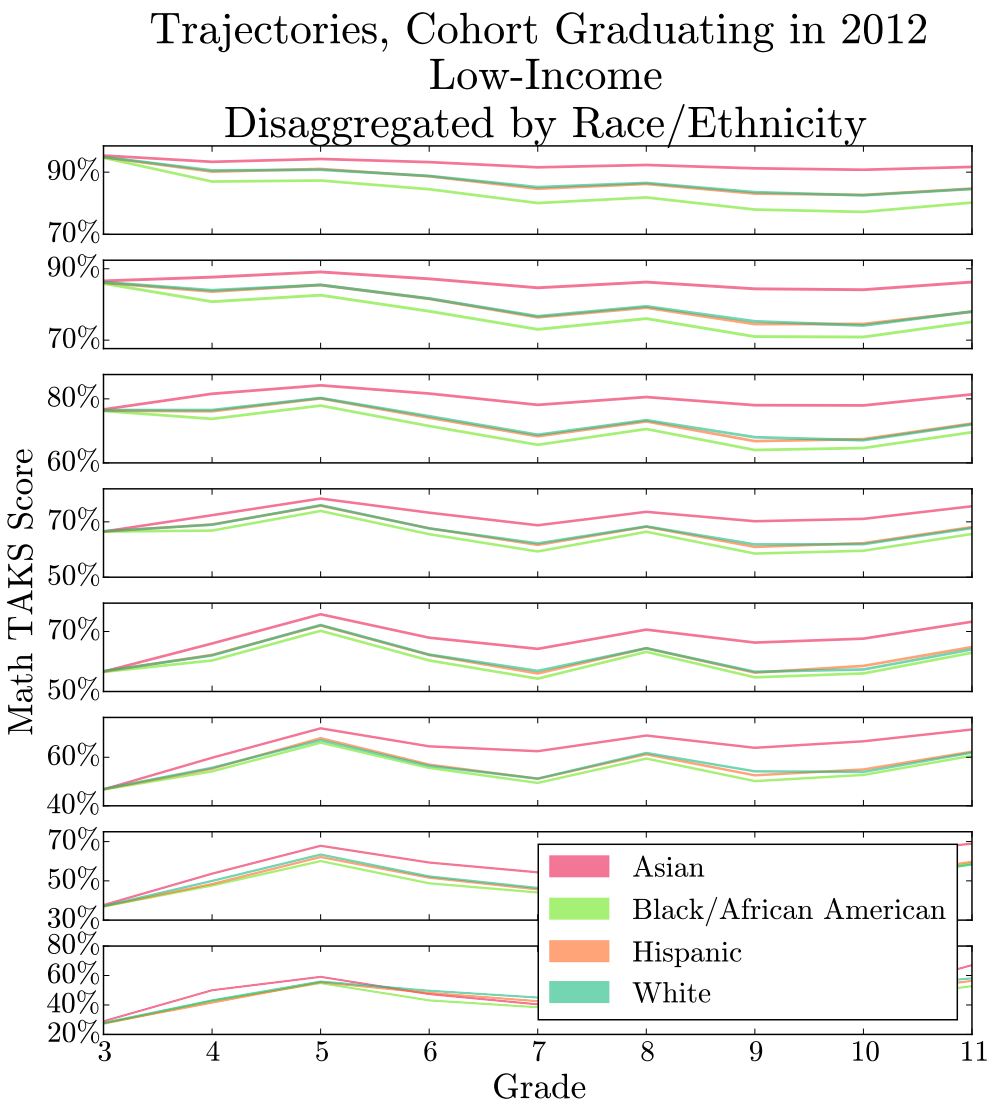

Figure 4.11: Trajectories for the low-income students in the cohort of 2012, disaggregated by race/ethnicity. Despite having a similar SES, performance disparities still exist.

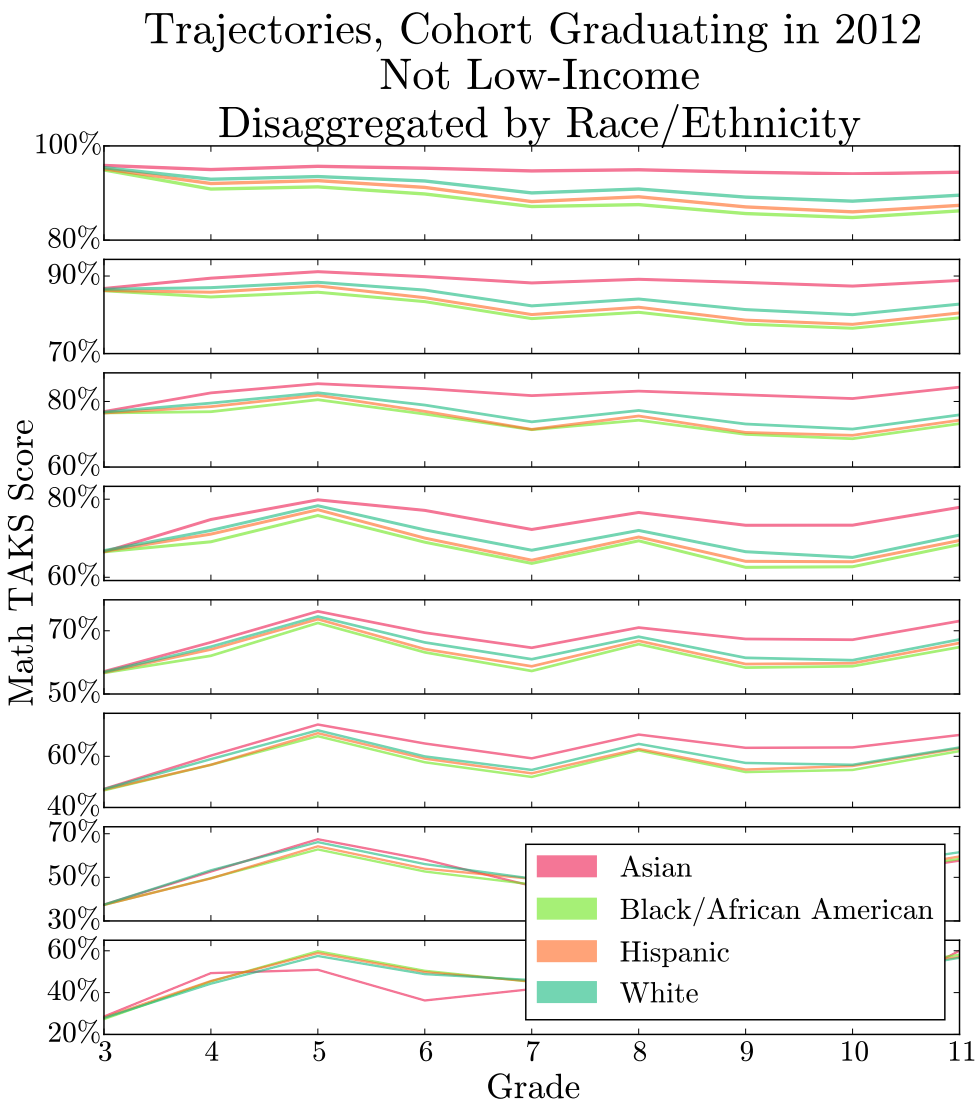

Figure 4.12: Trajectories for the not low-income students in the cohort of 2012, disaggregated by race/ethnicity. Despite having a similar SES, performance disparities still exist.

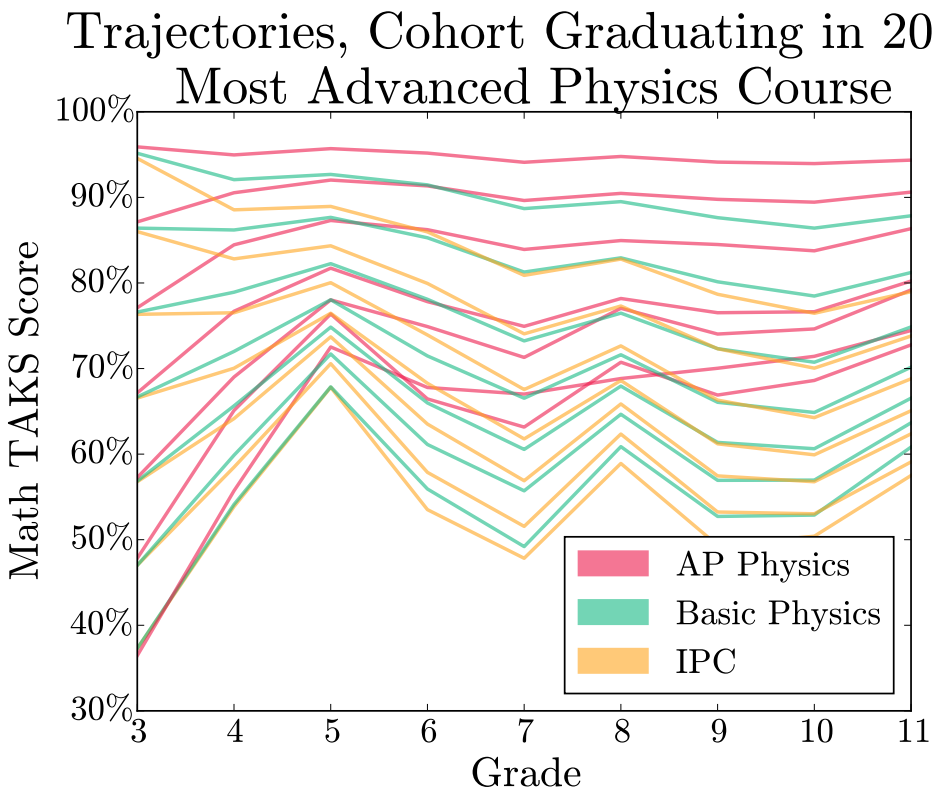

Figure 4.13: The trajectories for the cohort of 2012 disaggregated by highest level of physics course taking. Despite having similar 3rd grade scores, AP physics students outperform their basic physics and IPC counterparts.

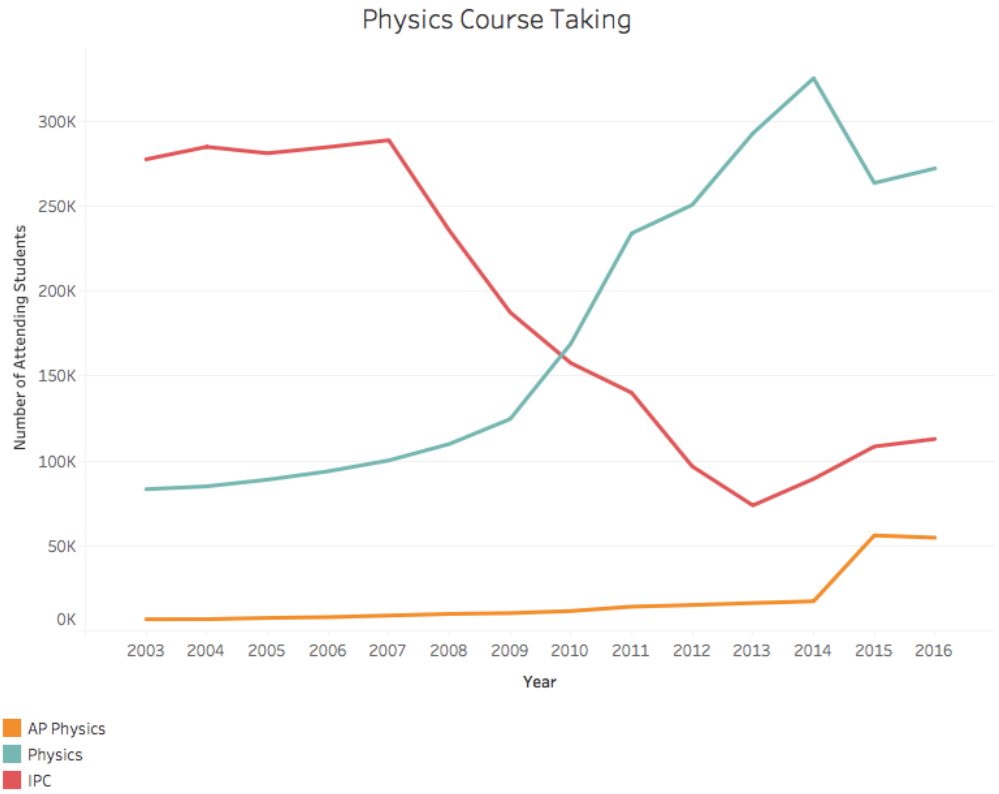

Figure 4.14: Numbers of students attending AP Physics, Physics, and IPC by year. 4x4 caused a decline in IPC enrollment and an increase in basic physics enrollment. HB-5 caused a reversal of this trend.

Chapter 5

Conclusions

Building off of the foundation created by Bansal, Bendinelli, and Marder [7, 8], I have developed an intuitive and accurate method, called AB snapshot streamlines, to predict longitudinal test scores using only 2 or 3 years of data. I use an alternative binning (AB) process to sort the students into score bins, determined by the raw percent score. Changes in score are calculated between two *other* exams, resulting in a vector field of score changes for each grade and score bin. Streamlines are interpolated between the arrows to create a continuous flow of student scores throughout the grades.

The AB snapshot streamlines utilize an accelerated longitudinal design by piecing together students from multiple overlapping cohorts to cover the full grade range of 3-11 in only 2 or 3 years. Trajectories and AB cohort streamlines, by comparison, use data from a single cohort, requiring nine years of data to cover the same grade range. These techniques developed naturally from our fluid mechanics backgrounds although certain aspects resemble techniques already implemented in various statistical fields. In particular, elements of our techniques are used in age-period-cohort studies and in group-based trajectory modeling [52, 73].

AB snapshot streamlines are unique in that they are graphical and non-parametric. As a result, AB snapshot streamlines are easy to interpret and require few assumptions. The technique is applicable in many fields, and its use has been demonstrated in this dissertation with standardized testing data. In particular, I investigated the effects of the Student Success Initiative, a program that provided both assistance and consequences for failing students. I hope that this technique can be used more broadly to answer research questions that require a more flexible and intuitive analysis approach.

Addressing regression to the mean and the random fluctuations in educational data is important and urgent. A new school and district accountability system that was passed in May of this year will be used consequence-free in the 2018-2019 school year and will take full effect in the 2019-2020 school year. House Bill 22 mandated that schools and districts be rated on an A-F scale in five domains: student achievement, student progress, closing performance gaps, postsecondary readiness, and community and student engagement. For the student achievement measure, the TEA will classify students each year by using cut-off scores. The random fluctuations in the observed scores can cause students to be misclassified, possibly pushing the school into a different performance rating. The student progress measure is most affected by regression to the mean, because points are rewarded to the schools for students who move up a classification but not for students who move down. The inevitable regression to the mean would impact schools with varying score distributions differently. A report by the TEA states that "the agency has begun examining several

alternative approaches to ensure we have the most effective method for recognizing student growth, but at present, no changes have been proposed" [98]. Analysis using the AB method could improve the new accountability system and have a profound impact on schools and districts in Texas.

Appendix

0.1 Z-score Properties

Claim: For z-scores—defined as $z_i = (x_i - \mu)/\sigma_{x_i}$, where σ_{x_i} is the standard deviation of raw score x_i and μ is the mean raw score or expected score $E(x_i) = \mu$—the expectation value and standard deviation are given by $E(z_i) = 0$ and $\sigma_{z_i} = 1$ for all i.

Proof:

$$E(z_i) = E((x_i - \mu)/\sigma_{x_i}) = \frac{1}{\sigma_{x_i}}(E(x_i) - \mu) = \frac{1}{\sigma_{x_i}}(\mu - \mu) = 0$$
$$\sigma_{z_i} = \sigma(x_i - \mu/\sigma_{x_i}) = \sigma_{x_i}/\sigma_{x_i} = 1$$

0.2 Pearson Correlation Coefficient Magnitude

Claim: For Pearson correlation coefficients—defined as $\rho_{x,y} = \sigma_{x,y}/(\sigma_x\sigma_y)$—the absolute value is never greater than 1, $|\rho_{z_i,z_j}| \leq 1$.

Proof: From the Cauchy-Schwarz inequality, $|E(\theta\phi)|^2 \leq E(\theta^2)E(\phi^2)$. If we set $\theta = x - E(x)$ and $\phi = y - E(y)$ then

$$|E((x - E(x))(y - E(y)))|^2 \leq E((x - E(x))^2)E((y - E(y))^2)$$
$$|E(xy) - E(x)E(y) - E(x)E(y) + E(x)E(y)|^2 \leq$$
$$(E(x^2) + E(x)^2 - 2E(x)^2)(E(y^2) + E(y)^2 - 2E(y)^2)$$
$$|E(xy) - E(x)E(y)|^2 \leq (E(x^2) - E(x)^2)(E(y^2) - E(y)^2)$$
$$|\sigma_{x,y}|^2 \leq \sigma_x^2\sigma_y^2$$
$$\implies \rho_{x,y}^2 \leq 1$$
$$\implies |\rho_{x,y}| \leq 1$$

0.3 Linear Conditional Expectation Value

Claim: If x and y are jointly normal distributions, the conditional expectation value will be linear: $E(y|x) = \alpha + \beta x$. Furthermore, the coefficients are determined such that $E(y|x) = E(y) + \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x - E(x)) = \mu_y + \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x - \mu_x)$.

Definitions:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left[\frac{-(x-\mu_x)^2}{2\sigma_x^2}\right]$$
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \exp\left[\frac{-\delta}{2(1-\rho_{x,y}^2)}\right]$$
$$\delta \equiv \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho_{x,y}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}$$
$$E(y|x) = \int y\frac{f(x,y)}{f(x)}dy$$

Proof:

$$\begin{aligned}\frac{f(x,y)}{f(x)} &= \frac{\sqrt{2\pi\sigma_x^2}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right.\right. \\ &\qquad \left.\left. - \frac{2\rho_{x,y}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right) + \frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \\ &= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho_{x,y}^2)}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(\frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho_{x,y}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right.\right. \\ &\qquad \left.\left. + \frac{\rho_{x,y}^2(x-\mu_x)^2}{\sigma_x^2}\right)\right] \\ &= \frac{1}{\sigma_y\sqrt{2\pi(1-\rho_{x,y}^2)}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(\frac{(y-\mu_y)}{\sigma_y} - \frac{\rho_{x,y}(x-\mu_x)}{\sigma_x}\right)^2\right]\end{aligned}$$

let $u = \frac{(y-\mu_y)}{\sigma_y}$, so $y = \mu_y + \sigma_y u$ and $dy = \sigma_y du$.

$$
\begin{aligned}
E(y|x) &= \int \frac{\mu_y + \sigma_y u}{\sqrt{2\pi(1-\rho_{x,y}^2)}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(u - \frac{\rho_{x,y}(x-\mu_x)}{\sigma_x}\right)^2\right] du \\
&= \int \frac{\mu_y}{\sqrt{2\pi(1-\rho_{x,y}^2)}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(u - \frac{\rho_{x,y}(x-\mu_x)}{\sigma_x}\right)^2\right] du \\
&+ \int \frac{\sigma_y u}{\sqrt{2\pi(1-\rho_{x,y}^2)}} \exp\left[\frac{-1}{2(1-\rho_{x,y}^2)}\left(u - \frac{\rho_{x,y}(x-\mu_x)}{\sigma_x}\right)^2\right] du \\
&= \mu_y + \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x-\mu_x)
\end{aligned}
$$

Bibliography

[1] Department of Assessment and Accountability. *2012 Adequate Yearly Progress (AYP) Guide*. Texas Education Agency, June 2012.

[2] Student Assessment Division. Technical Digest 2009-2010, 2010.

[3] Texas Education Agency. The Student Success Initiative: 2009-2010 Biennium Evaluation Report, 2011.

[4] Texas Legislative Budget Board. Fiscal Size-Up 2016-17 Biennium, May 2016.

[5] Texas Education Agency. State Graduation Requirements. http://tea.texas.gov/graduation.aspx.

[6] Kevin Grimm. Basics of Structural Equation Modeling and The Mplus Computer Program, 2008.

[7] M. Marder and D. Bansal. Flow and diffusion of high-stakes test scores. *Proceedings of the National Academy of Sciences*, 106(41):17267–17270, 2009.

[8] Anthony Bendinelli and Michael Marder. Visualization of Longitudinal Student Data. *Physical Review Special Topics - Physics Education Research*, 8(020119), 2012.

[9] Office of Technology Assessment U.S. Congress. *Testing in American Schools: Asking the Right Questions*, chapter Lessons From the Past: A History of Educational Testing in the United States. U.S. Government Printing Office, 1992.

[10] Dan Fletcher. Brief history: Standardized testing. *TIME*, December 11 2009.

[11] Janet Thomas and Kevin Brady. The Elementary and Secondary Education Act at 40: Equity, Accountability, and the Evolving Federal Role in Public Education. *Review of Research in Education*, 29:51–67, 2005.

[12] Elementary and Secondary Education Act of 1965. https://www.gpo.gov/fdsys/pkg/STATUTE-79/pdf/STATUTE-79-Pg27.pdf.

[13] House Resolution 610. https://www.congress.gov/bill/115th-congress/house-bill/610, 2017.

[14] Julie Hirschfeld Davis. President Obama Signs Into Law a Rewrite of No Child Left Behind. *The New York Times*, December 10 2015.

[15] Brian Stecher, Georges Vernez, and Paul Steinberg. Accountability for NCLB. https://www.rand.org/pubs/periodicals/rand-review/issues/summer2010/nclb.html, 2010.

[16] Texas Education Agency. AYP Results for Years 2003-2012 in Texas. https://rptsvr1.tea.texas.gov/ayp/2012/summaries12_exp.pdf, 2012.

[17] William Erphenbach. A Study of States' Requests for Waivers from Requirements of the No Child Left Behind Act of 2001, March 2014.

[18] Texas Education Agency. NCLB-ESEA Waiver Information. http://tea.texas.gov/Texas_Schools/Waivers/NCLB-ESEA_Waiver_Information/.

[19] U.S. Department of Education. No child left behind: A toolkit for teachers. https://www2.ed.gov/teachers/nclbguide/toolkit_pg6.html, 2009.

[20] Cory Koedel, Kata Mihaly, and Jonah Rockoff. Value-added modeling: A review. *Economics of Education Review*, 2015.

[21] Noelle Paufler and Audrey Amrein-Beardsley. The random assignment of students into elementary classrooms. *American Educational Research Journal*, 51(2):328–362, 2014.

[22] Eric Hanushek. The economic value of higher teacher quality. *Economics of Education Review*, 30(3):266–479, 2011.

[23] Raj Chetty, John Friedman, and Jonah Rockoff. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679, 2014.

[24] Jonah Rockoff, Douglas Staiger, Thomas Kane, and Eric Taylor. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, 102(7):3184–3213, 2012.

[25] Scott Condie, Lars Lefgren, and David Sims. Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40(1):76–92, 2014.

[26] Steven Glazerman, Ali Protik, Bing ru Teh, Julie Bruch, and Jeffrey Max. *Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment*. United States Department of Education, 2013.

[27] Los Angeles Times. Los Angeles Teacher Ratings. http://projects.latimes.com/value-added/.

[28] The University of Texas at Austin. Texas education research center. https://research.utexas.edu/erc/.

[29] Family Policy Compliance Office. Family Educational Rights and Privacy Act (FERPA). https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

[30] The University of Texas at Austin Education Research Center. *The University of Texas at Austin Education Research Center Policies and Procedures: General Information*. The University of Texas at Austin, October 2015.

[31] UTeach: We Prepare Teachers. They Change the World. https://uteach.utexas.edu/.

[32] Student Assessment Division. Technical Digest 2014-2015, 2015.

[33] Pearson Educational Measurement. TAKS Higher Education Readiness Component Contrasting Groups Study, 2006.

[34] Texas Education Code. Title 3, Subtitle A, Chapter 51: Provisions Generally Applicable to Higher Education.

[35] Texas Education Agency. Student Success Initiative Manual (2017). http://tea.texas.gov/student.assessment/SSI/.

[36] Texas Education Agency. The Student Success Initiative: An Evaluation Report (2009). tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147490898.

[37] Texas Classroom Teachers Association. More details on budget proposals. https://tcta.org/politics_government/updates_from_the_capitol/14407-more_details_on_budget_proposals, January 2017.

[38] Texas 85th Legislative Session. SB1 General Appropriations Bill. www.lbb.state.tx.us/Documents/Appropriations_Bills/85/Conference_Bills/SB1_Conference_Bill.pdf, 2017.

[39] Melvin R. Novick. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3:1–18, 1966.

[40] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publication Corporation, 1968.

[41] Lee Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 1951.

[42] Gregory Camilli. Origin of the Scaling Constant d=1.7 in Item Response Theory. *Journal of Educational and Behavioral Statistics*, 19(3):293–295, 1994.

[43] Frank Baker. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, 2001.

[44] Texas Education Code. Title 2, Subtitle H, Chapter 39: Public School System Accountability.

[45] Benjamin Wright, Ronald Mead, and Susan Bell. *BICAL: Calibrating Items with the Rasch Model*. University of Chicago, 1980.

[46] Lesa Hoffman. *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. Routledge, 2015.

[47] Stephen Raudenbush and Anthony Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, 2 edition, 2002.

[48] Judith Singer and John Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurence*. Oxford University Press, 2003.

[49] John McArdle and John R. Nesselroade. *Longitudinal Data Analysis Using Structural Equation Models*. American Psychological Association, 2014.

[50] Patrick Curran and Bengt Muthen. The application of latent curve analysis to testing developmental theories in intervention research. *American Journal of Community Psychology*, 27(4), 1999.

[51] Bengt Muthen and Siek-Toon Khoo. Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences*, 10(2):73–101, 1998.

[52] Daniel Nagin. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2):139–157, 1999.

[53] Heather Andruff, Natasha Carraro, Amanda Thompson, Patrick Gaudreau, and Benoit Louvet. Latent Class Growth Modelling: A Tutorial. *Tutorials in Quantitative Methods for Psychology*, 5(1):11–24, 2009.

[54] S. Raudenbush, A. Bryk, and R. Congdon. *HLM 7.01 for Windows [Computer Software]*. Scientific Software International Inc., Skokie, IL, 2013.

[55] L.K. Muthen and B.O. Muthen. *Mplus User's Guide*. Muthen and Muthen, Los Angeles, CA, 8 edition, 2017.

[56] Patrick Curran, Khawla Obeidat, and Diane Losardo. Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, 11(2):121–136, 2010.

[57] Bengt Muthen. Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), Handbook

of Quantitative Methodology for the Social Sciences. Newbury Park, CA: Sage Publications, 2004.

[58] Daniel Nagin and Richard Tremblay. Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, 6(1):18–34, 2001.

[59] Daniel Nagin, Bobby Jones, Valeria Passos, and Richard Tremblay. Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, 0(0):1–9, 2016.

[60] Lawrence Kupper, Joseph Janis, Azza Karmous, and Bernard Greenberg. Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases*, 38(10):811–830, 1985.

[61] Robert Robinson and Elton Jackson. Is Trust in Others Declining in America? An Age-Period-Cohort Analysis. *Social Science Research*, 30:117–145, March 2001.

[62] April Clark and Marie Eisenstein. Interpersonal trust: An age-period-cohort analysis revisited. *Social Science Research*, 42(2):361–375, March 2013.

[63] Yang Yang and Kenneth Land. Age-period-cohort analysis of repeated cross-section surveys. *Sociological Methods and Research*, 36(3):297–326, February 2008.

[64] Marie Dahlin, Nils Joneborg, and Bo Runeson. Stress and depression among medical students: A cross-sectional study. *Medical Education*, 39:594–604, 2005.

[65] Vittorio Busato, Frans Prins, Jan Elshout, and Christiaan Hamaker. Learning styles: a cross-sectional and longitudinal study in higher education. *British Journal of Educational Psychology*, 68:427–441, 1998.

[66] Charlene Krueger and Lili Tian. A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research for Nursing*, 6(2):151–157, October 2004.

[67] Norman B. Ryder. The cohort as a concept in the study of social change. *American Sociological Review*, 30(6):843–861, December 1965.

[68] Carrie Conaway, Venessa Keesler, and Nathaniel Schwartz. What Research Do State Education Agencies Really Need? The Promise and Limitations of State Longitudinal Data Systems. *Educational Evaluation and Policy Analysis*, 37(1S):16S–28S, May 2015.

[69] Texas Education Agency. House Bill 3 Transition Plan. http://tea.texas.gov/student.assessment/hb3plan/.

[70] Texas Education Agency. Policy Changes and the Graduation Rate. http:// tea.texas.gov/WorkArea/DownloadAsset.aspx?id=2147483983.

[71] National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future.* The National Academies Press, 2007.

[72] Richard Bell. Convergence: An Accelerated Longitudinal Approach. *Child Development*, 24(2):145–152, June 1953.

[73] Yasuo Miyazaki and Stephen Raudenbush. Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, 5(1):44–63, 2000.

[74] Sally Galbraith, Jack Bowden, and Adrian Mander. Accelerated longitudinal designs: An overview of modelling, power, costs, and handling missing data. *Statistical Methods in Medical Research*, 26(1):374–398, 2017.

[75] Mirjam Moerbeek. The effects of the number of cohorts, degree of overlap among cohorts, and frequency of observation on power in accelerated longitudinal designs. *Methodology*, 7:11–24, 2011.

[76] John Mirowsky and Jinyoung Kim. Graphing age trajectories: Vector graphs, synthetic and virtual cohort projections, and cross-sectional profiles of depression. *Sociological Methods and Research*, 35(4):497–541, May 2007.

[77] John Mirowsky. Depression and the sense of control: Aging vectors, trajectories, and trends. *Journal of Health and Social Behavior*, 54(4), 2013.

[78] Anthony Bendinelli. *The Application of Visualization Methods to Educational Data Sets with Inspiration from Statistical and Fluid Mechanics*. PhD thesis, University of Texas at Austin, May 2014.

[79] Y. Nakayama. *Introduction to Fluid Mechanics*. Butterworth-Heinemann, 1999.

[80] Teddy Schall and Gary Smith. Do baseball players regress toward the mean? *The American Statistician*, 54(4):231–235, November 2000.

[81] Lant Pritchett and Lawrence Summers. Asiaphoria meets regression to the mean. *National Bureau of Economic Research*, 2014.

[82] Gary Smith and Joanna Smith. Regression to the mean in average test scores. *Educational Assessment*, 10(4), 2005.

[83] John R. Nesselroade, Stephen M. Stigler, and Paul B. Baltes. Regression toward the mean and the study of change. *Psychological Bulletin*, 88(3):622–637, 1980.

[84] Stephen M. Stigler. Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, 6:103–114, 1997.

[85] HM Lin and MD Hughes. Adjusting for regression toward the mean when variables are normally distributed. *Statistical Methods in Medical Research*, 6:129–146, 1997.

[86] Christy Chuang-Stein and Donald Tong. The impact and implication of regression to the mean on the design and analysis of medical investigations. *Statistical Methods in Medical Research*, 6:115–128, 1997.

[87] Garrett Fitzmaurice. Regression to the mean. *Nutrition*, 16:81–82, 2000.

[88] Wold Schwarz and Dennis Reike. Regression away from the mean: Theory and examples. *British Journal of Mathematical and Statistical Psychology*, 2017.

[89] T. L. Kelley. *Fundamentals of Statistics*. Harvard University, Cambridge, MA, 1947.

[90] J Martin Bland and Douglas G Altman. Some examples of regression towards the mean. *BMJ*, 309:780, 1994.

[91] Peter Rousseeuw. Why the wrong papers get published. *Chance: New Directions for Statistics and Computing*, 4(1), 1991.

[92] Horace Secrist. *The triumph of mediocrity in business*. Bureau of Business Research, Northwestern University, Evanston, IL, 1933.

[93] Michael Maraun, Stephanie Gabriel, and Jack Martin. Thy mythologization of regression towards the mean. *Theory and Psychology*, 21(6):762–784, 2011.

[94] Andrew Chesher. Non-normal variation and regression to the mean. *Statistical Methods in Medical Research*, 6:147–166, 1997.

[95] Simon Tidd and Sonia Dominguez. Chronic absenteeism and student mobility. http://e3alliance.org/wp-content/ uploads/2017/04/E3-3D-Mobility-Chronic-Absence-040417.pdf, 2017.

[96] Jeffrey Weiss. Three things you should know about the 2015 STAAR results. *The Dallas Morning News*, May 2015.

[97] UTexas Admissions. Coursework requirements. https://admissions.utexas.edu/explore/prerequisites/general-requirements.

[98] Texas Education Agency. *2015-16 A-F Ratings*, December 2016.

Vita

Sarah Emily Elias Stephens (formerly Sarah Emily Elias) received her Bachelor of Science degree in Physics and Mathematics from Brandeis University in Waltham, Massachusetts in 2010. She immediately started her graduate studies at the University of Texas at Austin. She worked with co-advisors Xiaoqin Li and Andrea Alù until 2014. She began working with Michael Marder in early 2015.

Permanent address: saraheestephens@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.